Measuring school performance for early elementary grades in Maryland

Lisa Dragoset, Cassandra Baxter, Dallas Dotter, Elias Walsh

December 2019

REL
MID-ATLANTIC
Regional Educational Laboratory

Measuring school performance for early elementary grades in Maryland

Lisa Dragoset, Cassandra Baxter, Dallas Dotter, Elias Walsh

December 2019

Key findings

A K–3 school-level growth measure was estimated and examined. The study identified some concerns about its validity and precision that suggest it should be used for accountability with caution.

- The overall Kindergarten Readiness Assessment score performs about as well in predicting grade 3 achievement as combinations of kindergarten readiness subscores.
- Schools' K–3 growth estimates are likely less valid than schools' grade 3–4 growth estimates but have a similar level of precision.
- Schools' K–3 growth estimates are much less precise for smaller schools than for larger schools.
- Administering the Kindergarten Readiness Assessment to a subset of students in each classroom (as opposed to all students) greatly reduces the precision of schools' K–3 growth estimates.

Why this study?

The Maryland State Department of Education (MSDE) has a critical need to better understand its schools' contributions to student learning in the early elementary grades as a part of its accountability system under the Every Student Succeeds Act (ESSA). The early grades lay an important foundation for students' future academic success. Yet, as is the case for nearly every state, MSDE lacks a measure of how well its schools are supporting the academic growth of its youngest students, from kindergarten to grade 3.

Growth measures, which estimate schools' contributions to students' assessment scores, are a critical component of accountability systems, and the absence of a growth measure for the early grades can have a detrimental impact on early learning. Growth measures, and accountability systems more broadly, inform how resources, policies, and practices can be adjusted to better support student learning by identifying areas of strength or need that warrant further investigation. Growth measures are widely used across states in late elementary and middle grades, but are lacking for early grades, though other measures could be used for accountability in those grades, such as chronic absenteeism and school climate. States' abilities to effectively guide policy are hindered when there is no information about how well their schools and teachers are promoting student learning in the critical early grades.

MSDE sought to investigate the feasibility of expanding its current accountability system to include a growth measure from kindergarten to grade 3. Growth is one of four indicators that Maryland uses in its accountability system for elementary and middle schools and is worth one-quarter of a school's accountability points. Currently, growth for elementary schools is measured only for students in grades 4 and 5 (Maryland State Department of Education, 2018). Having a growth measure from kindergarten to grade 3 would enable MSDE to hold elementary schools accountable for student growth in all grades, as it does for middle schools, and help inform policies aimed at improving early learning.

Maryland's current growth measures begin in grade 4 because its primary statewide assessment—the Partnership for Assessment of Readiness for College and Careers (PARCC)—is first administered in grade 3, which provides a baseline for measuring growth. MSDE uses PARCC assessments to measure schools' contributions to student learning, or growth, in reading and in math, annually between grades 3 and 8, using student growth percentiles (SGPs). These measures are used to hold schools accountable for how their students' assessment scores, within the same subject, change from one year to the next.

In 2014, MSDE launched its Kindergarten Readiness Assessment (KRA) and administered it to every kindergartener at the start of the 2014/15 school year, establishing a baseline measure of student readiness at the point of school entry. The purpose of the KRA is to provide information about how well prepared children are for kindergarten, which enables programmatic decision-making at the school, district, and state levels. In the 2017/18 school year, this first KRA cohort completed the grade 3 PARCC, which created an opportunity to measure growth from kindergarten entry through the end of grade 3, or K–3, for the first time.

MSDE partnered with the Regional Educational Laboratory (REL) Mid-Atlantic to examine whether it was feasible to construct a K–3 SGP growth measure that could be used for accountability purposes. Depending on the feasibility of developing a school-level measure from these two assessments, MSDE will examine whether and how it should include the measure to align with its current points system for ranking schools in its accountability system.

Incorporating a K–3 measure in Maryland's accountability system would break new ground. No other state currently measures K–3 student growth statewide for accountability purposes (O'Keefe, 2017) and few states have measured growth using a kindergarten assessment, which typically differ from later grade assessments in format and scope. Equipped with a valid K–3 growth measure, MSDE would be able to identify elementary schools with low- and high-performing early grades, and more effectively guide policy and resources to improve those schools. With more than 25 states using kindergarten entry assessments (REL Northwest, 2017), this study provides a blueprint for constructing early elementary growth measures using different assessments that are administered more than one year apart.

Research questions

This study explored four primary research questions related to whether a school-level K–3 growth measure could be developed for accountability purposes in Maryland:

1. Does the growth model perform as well with the overall KRA score compared to KRA subscores?

As is true in most states, Maryland's assessments administered in kindergarten and grade 3 are distinct and have different properties and goals. While the PARCC measures students' performance on grade 3 standards for reading and math, the KRA assesses students' kindergarten readiness in four domains—language and literacy, mathematics, social foundations, and physical well-being and motor development. On the KRA, students receive an overall scaled score and a scaled subscore for each of the four domains (see table A.2 in appendix A for more information on the PARCC and KRA assessments and scores). Some of these subscores may relate more closely to grade 3 performance than others, and growth estimates will be most valid if they use a configuration of the KRA score that most closely predicts grade 3 PARCC scores. This study examined which KRA score was best suited as a baseline measure in the growth model: the KRA overall score, certain domain subscores, or a specific combination of domain subscores.

2. Are schools' K–3 growth estimates valid and precise, relative to the estimates used for accountability in later grades?

__Validity__: Is the growth estimate credible? Does the growth estimate appear to be measuring what it is intended to measure: schools' true contributions to their students' K–3 growth? That is, is student academic performance, as measured by the assessments used in the model, related?

__Precision__: Is the growth estimate a consistent measure? That is, will schools' estimates vary from year to year even if their true performance is not changing?

In Maryland, students with high KRA scores tended to also have high grade 3 PARCC scores (see figures B.1 and B.2 in appendix B). However, to produce *valid* estimates of K–3 growth, performance on the KRA must capture aspects of student academic ability that benefit from K–3 instruction and are measured by performance on the grade 3 PARCC. If the relationship between KRA and grade 3 PARCC scores is weaker than the relationships between PARCC scores for different grade levels, it would suggest that the KRA and grade 3 PARCC are measuring different aspects of academic ability, potentially compromising the measure's ability to accurately measure schools' contributions to student academic growth. If estimates are not sufficiently precise, it will be difficult to determine whether changes in schools' estimates reflect true changes in performance or noise in the data.

3. How does school size affect precision of K-3 growth estimates?

Growth estimates for smaller schools are typically less precise than those for larger schools, but the loss of precision may differ for different assessments and settings. This study examined how the precision of schools' K–3 growth estimates in Maryland change relative to schools' size, and the extent to which growth estimates are less precise for smaller schools than for larger schools.

4. How would administering the kindergarten assessment to a random subsample of students affect the precision of the growth estimate?

Kindergarten assessments that are individually administered by a teacher to each student, like the KRA, can be costly to implement. A 2016 Maryland law allowed local school systems to administer the KRA to a random subsample of kindergarteners (a "partial-cohort administration"), rather than to all kindergarteners, beginning with the 2016/17 school year. This study examined how the precision of the estimates will be affected by the partial-cohort administration. This information will be useful to other states as they weigh the costs and benefits of sampling approaches to administering assessments.

The data and methods used to explore these research questions are described in Box 1 and appendix A.

Box 1. Data sources, sample, and methods

Data sources. The study used administrative data provided by the Maryland State Department of Education (MSDE). Assessment score data included Kindergarten Readiness Assessment (KRA) scaled scores from the 2014/15 school year and grades 3–6 Partnership for Assessment of Readiness for College and Careers (PARCC) reading and math scaled scores from 2014/15 through 2017/18. The data also included student demographics and attendance data, for the school that the student was enrolled in on the last day of the school year, for students in grades K–6 in the 2014/15 to 2017/18 school years. Data were linked across files using student identification codes. A complete list of the data sources are included in appendix A.

Sample. Students who had a valid 2014/15 KRA score and a valid 2017/18 grade 3 PARCC score were included in all analyses. A total of 54,393 students were included in the math model and 54,397 students were included in the reading model (these students represent 86 percent of all students with a 2014/15 KRA score; see appendix A for more information). Students with non-traditional grade progression were excluded from the sample. Students with significant cognitive disabilities (who took the alternate assessment) were also excluded. To understand the statistical properties of the K–3 Student Growth Percentile (SGP) estimates, research question 2 and supplemental analyses that are described in appendix A drew on an additional sample of 359,619 students with scores on any of the 2014/15 to 2016/17 grade 3, 2014/15 to 2017/18 grades 4 and 5, or 2015/16 to 2017/18 grade 6 PARCCs (see appendix A for specific cohort and year combinations used in these analyses). Appendix A includes a complete list of business rules for defining the sample.

Methodology

The SGP model. The SGP model used overall KRA scaled scores to group students into peer groups based on their academic performance at kindergarten entry and then assessed the student's current performance in grade 3, as measured by PARCC scaled scores, relative to their group of academic peers at kindergarten entry. SGPs were estimated separately for grade 3 reading and math, and then SGPs were aggregated to the school level by calculating the mean SGP (mSGP) among the students who attended the school, providing a measure of the school's contributions to a typical student's academic growth. The SGP model accounts for measurement error (the extent to which scores do not reflect students' actual ability) in KRA scores.

Research question 1. Ordinary least squares regression and pairwise correlations were used to determine how well each of four versions of the students' KRA score predicted their grade 3 PARCC scores. Pairwise correlations were interpreted throughout the report using the following classifications: weak (0.1-0.39), moderate (0.4-0.69), strong (0.7-0.99), and perfect (1) (Dancey and Reidy 2007). Differences between correlations were evaluated based on statistical significance testing.

Because minor differences can be statistically significant when using large samples, differences were also assessed for practical meaning using the weak/moderate/strong/perfect classifications noted above. The results of these analyses were used to determine which KRA score would be used to calculate growth estimates for the remaining research questions.

Research question 2. To examine the validity of a K-3 growth estimates, the study team used pairwise correlations to assess the strength of the relationship between students' KRA and grade 3 PARCC scores, relative to relationships between students' grades 3 and 4, and grades 3 and 6 PARCC scores.

*Definition of a 95 percent confidence interval****: This interval is a range of values that would contain the school's true mean SGP 95 percent of the time, if the SGP estimation was repeated many times using different random samples of students in the school.***

To examine precision, the study team calculated a 95 percent confidence interval around each school's growth estimate. A wider average confidence interval indicates less precision. The study team compared the average confidence interval for schools' K–3 and grades 3–4 growth estimates (calculated using the SGP model described above) to assess the precision of the K–3 estimate relative to an existing growth measure in Maryland's accountability system.

Research question 3. To help MSDE determine whether to report growth estimates for small schools, the study team assessed the width of 95 percent confidence intervals for growth estimates in relation to number of the school's students who took both exams. Confidence interval widths were defined as substantially different here and in research question 4 if they differed by at least 50 percent.

Research question 4. To estimate the potential impact of randomly sampling students to complete the KRA on precision, the study team recalculated schools' 2014/15 K–3 growth estimates using random samples of students that mimicked how partial-cohort administration was conducted in future years. The width of the confidence intervals around these estimates were then compared to those of the K–3 estimates using the full sample.

More detailed information on the methodologies is provided in appendix A.

Findings

The overall Kindergarten Readiness Assessment score performs about as well in predicting grade 3 achievement as combinations of kindergarten readiness subscores

This study evaluated how well each of four different configurations of KRA scores predicted students' grade 3 PARCC scores in math and reading. The four configurations were (1) the KRA overall score; (2) the KRA domain subscore that aligns with the SGP subject (that is, the math domain subscore for the math SGP and the reading domain subscore the for reading SGP); (3) a weighted combination of the KRA math and reading domain subscores; and (4) a weighted combination of all four KRA domain subscores, which gives larger weights to subscores that better predict grade 3 performance (see appendix A for details on how the weighted scores were obtained). The correlations between students' grade 3 PARCC scores and each of these KRA scores are presented in Table 1.

Table 1. Correlations between students' Kindergarten Readiness Assessment (KRA) scores and grade 3 Partnership for Assessment of Readiness for College and Careers (PARCC) scores

	Correlations between 2014/15 KRA scores and 2017/18 grade 3 PARCC scores	
	Math	Reading
Version 1: Overall scaled score	0.53	0.53
Version 2: Same-subject domain score (that is, math or reading)	0.53	0.48
Version 3: Weighted average of math and reading domain scores	0.55	0.54
Version 4: Weighted average of all domain scores	0.56	0.55

Note: See appendix A for details on how the weighted scores were obtained. For each subject (math and reading), the study tested whether (1) the correlation between version 1 (that is, KRA overall scaled score, which is shown in the first row of the table) differed from the other versions of the KRA score. All of these tests were significant at the 5 percent significance level, except the difference between the version 1 and version 2 for math.
Source**:** Administrative data provided by the Maryland State Department of Education.

The correlations between each configuration of the KRA score and grade 3 scores were not substantially different. For both math and reading, the weighted average of all the domain scores had the strongest relationship with grade 3 PARCC scores by a small margin (0.56 for math and 0.55 for reading), followed by the weighted average of the math and reading domain scores (0.55 for math and 0.54 for reading), the overall scaled score (0.53 for math and reading), and then the same subject domain score (0.53 for math and 0.48 for reading).

The results presented in table 1 suggest that the overall scaled score is likely to predict grade 3 performance about as well as the weighted scores. Additionally, the overall scaled score has the advantage of being a more straightforward measure that will be easier to communicate to educators and parents and easier to replicate in future years. The level of effort required to replicate a weighted score in future years, and explain to stakeholders how and why the score is changing, likely outweighs the marginal improvement such a score yields. Therefore, the study used the KRA overall scaled score when calculating the K–3 growth estimates that were examined in research questions 2 through 4.

Schools' K–3 growth estimates are likely less valid than schools' grades 3–4 growth estimates but have a similar level of precision

The study compared the strength of the relationship between students' KRA scores and grade 3 PARCC scores to the strength of the relationships between students' grade 3 and grade 4 PARCC scores for three cohorts of students, and between students' grade 3 and grade 6 PARCC scores (which involve a similar amount of time between assessments).

The study found that schools' K–3 growth estimates are likely less valid than schools' grades 3–4 growth estimates because the correlation between students' KRA and grade 3 scores is significantly lower than the correlation between students' grades 3 and 4 scores (table 2). The correlation between students' KRA and grade 3 scores is also significantly lower than the correlation between students' grade 3 and 6 PARCC scores.

Table 2. Correlation between students' initial and subsequent assessment scores, by cohort

	Correlation coefficient	
Grades and school years	Math	Reading
Between K (2014/15) and grade 3 (2017/18) scores	0.53	0.53
Between grade 3 (2014/15) and grade 4 (2015/16)	0.86	0.82
Between grade 3 (2015/16) and grade 4 (2016/17)	0.87	0.84
Between grade 3 (2016/17) and grade 4 (2017/18)	0.87	0.85
Between grade 3 (2014/15) and grade 6 (2017/18)	0.82	0.77

Note: For each subject (math and reading), the study tested whether (1) the correlation between kindergarten (KRA) and grade 3 (PARCC) scores (shown in the first row of the table) differed from the correlation between grade 3 and grade 4 PARCC scores (this test was run separately for each of the grade 3–4 cohorts, shown in the second through fourth rows of the table); (2) the correlation between grade 3 and grade 4 PARCC scores differed from the correlation between grade 3 and grade 6 PARCC scores shown in the last row of the table (this test was run separately for each of the grade 3–4 cohorts, shown in the second through fourth rows of the table); and (3) the correlation between kindergarten and grade 3 scores differed from the correlation between grade 3 and 6 scores. All of these tests were significant at the 5 percent significance level.
Source: Administrative data provided by the Maryland State Department of Education.

The study found that schools' K–3 growth estimates have a similar level of precision as schools' grades 3–4 growth estimates. The average confidence interval width for schools' K–3 growth estimates was 12 percentile points for math and 13 for reading, compared to 13 for schools' grades 3–4 growth estimates in both math and reading (using the grades 3–4 SGP estimates and confidence intervals that were calculated for this study; table 3). Precision is driven largely by sample size, that is, the number of students in a school. Thus, it is perhaps not surprising that K–3 and grades 3–4 growth estimates have similar levels of precision, as they are based on similar numbers of students.

Table 3. Average confidence interval width for schools' growth estimates, by cohort

	Average confidence interval width (percentile points)	
Grades and school years	Math	Reading
Between K (2014/15) and grade 3 (2017/18)	12	13
Between grade 3 (2014/15) and grade 4 (2015/16)	13	13
Between grade 3 (2015/16) and grade 4 (2016/17)	13	13
Between grade 3 (2016/17) and grade 4 (2017/18)	13	13

Note: See appendix A for details on how confidence intervals were calculated. Note that the study calculated schools' grades 3–4 SGP estimates and confidence intervals using the methods described in Appendix A, and these differ from the official measures and calculations used for accountability purposes in Maryland.

Source: Administrative data provided by the Maryland State Department of Education.

Schools' K–3 growth estimates are much less precise for smaller schools than for larger schools

Schools' K–3 growth estimates are much less precise for smaller schools than for larger schools. For example, for math, the average confidence interval width is 32 percentile points for schools with fewer than 15 tested students. A confidence interval of 32 percentile points means that an average school (with an estimated mean SGP of 50) could not be identified as different from a school with an estimated mean SGP of 34 (16 percentile points below 50) or 66 (16 percentile points above 50). In contrast, the average confidence interval for math was 8 percentile points for the largest schools (those with 180 to 208 tested students; table 4). All schools with 25 or more students have confidence interval widths less than 30 percentile points, and all schools with 50 or more students have confidence interval widths less than 20 percentile points.

Table 4. Average confidence interval width for schools' K–3 growth estimates, by school size

School size (percentile/range)[a]	Average confidence interval width for schools of that size[b]
Math	
1st/1–14 students	32
5th/15–30 students	19
10th/31–41 students	16
25th/42–59 students	14
50th/60–79 students	12
75th/80–107 students	11
90th/108–132 students	10
95th/133–148 students	9
99th/149–179 students	9
100th/180–208 students	8
Reading	
1st/1–14 students	27
5th/15–30 students	20
10th/31–41 students	17
25th/42–58 students	15
50th/59–80 students	13
75th/81–107 students	11
90th/108–132 students	10
95th/133–148 students	10
99th/149–179 students	9
100th/180–209 students	8

Note: See appendix A for details on how confidence intervals were calculated.

a. This column shows school size measured as the number of students contributing information to the school's K–3 growth estimate.

b. This column shows the average confidence interval width for schools in a particular quantile of school size. For example, the first row of the table shows the average confidence interval width for schools in the 1st percentile of school size (that is, schools that have 14 or fewer students contributing information to their growth estimate).

Source: Administrative data provided by the Maryland State Department of Education, 2014/15 to 2017/18.

Administering the Kindergarten Readiness Assessment to a subset of students in each classroom greatly reduces the precision of schools' K–3 growth estimates

As described above, beginning in school year 2016/17, Maryland law allowed districts to administer the KRA to a random subset of students in each classroom. In 2016/17, 2017/18, and 2018/19, 16, 12, and 10 districts (out of 24) chose this option (including all of the largest districts in the state, except for Baltimore City); the remaining districts administered the KRA to all students in the district. Statewide, 34 percent of students took the KRA in 2016/17, 35 percent took it in 2017/18, and 39 percent took it in 2018/19.

Administering the KRA to a subset of students in each classroom greatly reduces the precision of schools' K–3 growth estimates. The average width of confidence intervals around schools' K–3 growth estimates double from roughly 12 to roughly 25 percentile points (table 5). When all students take the KRA, the vast majority of schools have a confidence interval width less than 20 percentile points (figure 1 shows results for math using the 2016/17 sampling percentages; results for reading and for other sampling percentages are similar and are shown in appendix B). In contrast, when only a third of students take the KRA, more than half of schools have a confidence interval width greater than 20. Note that these estimates relate only to precision. Random sampling is unlikely to affect validity of growth estimates because the smaller random samples of students will yield the same estimates of growth on average, though with greater variability than using all students.

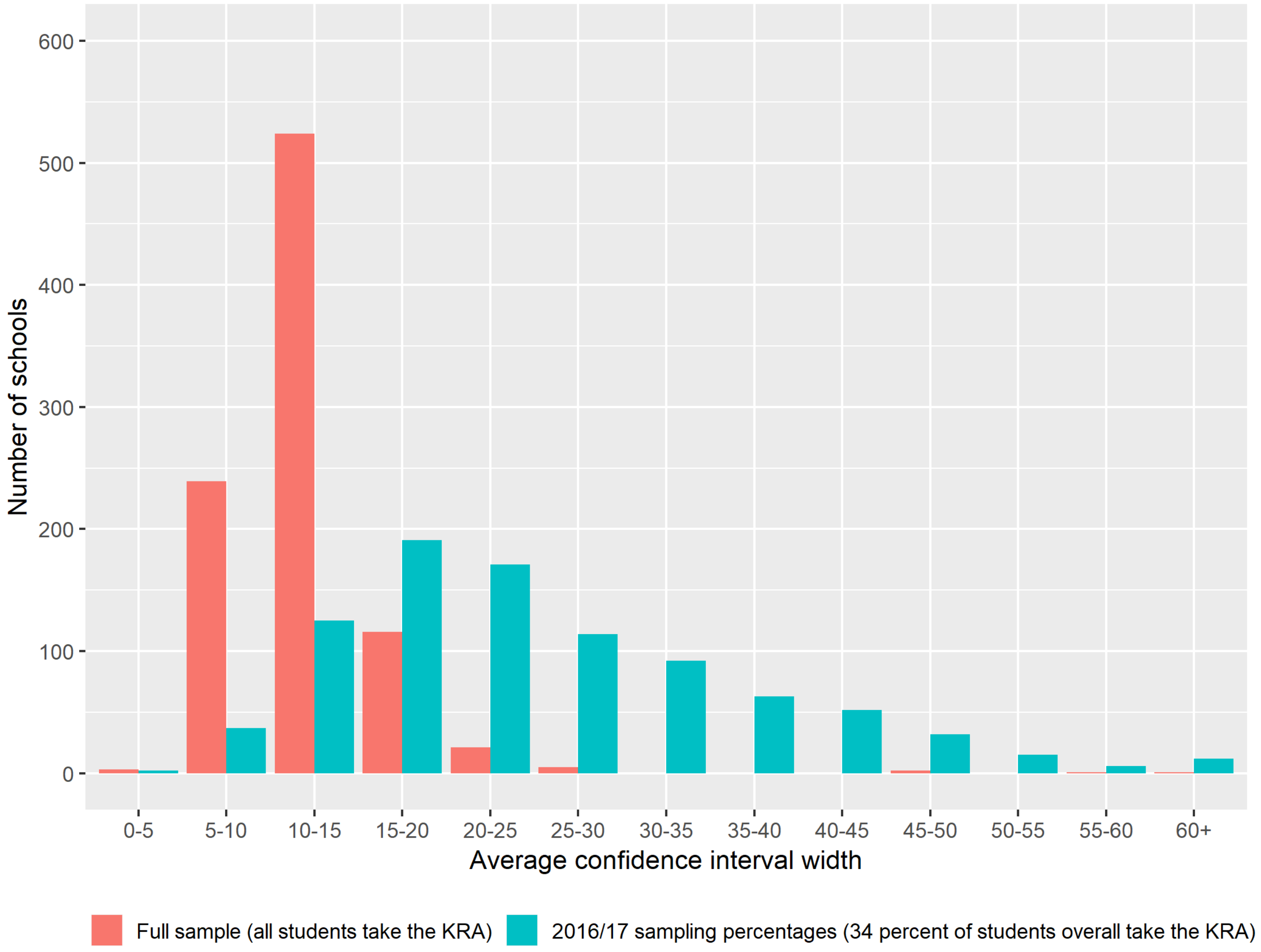
Table 5. Average confidence interval width for schools' K–3 growth estimates, by percentage of students who take the Kindergarten Readiness Assessment (KRA)

	Average confidence interval width (percentile points)	
Sampling percentage	Math	Reading
2014/15 full sample (all students take the KRA)	12	13
2016/17 sampling percentages (34 percent of students overall take the KRA)	25	26
2017/18 sampling percentages (35 percent of students overall take the KRA)	24	24
2018/19 sampling percentages (39 percent of students overall take the KRA)	23	23

Note: See appendix A for details on how confidence intervals were calculated.
Source: Administrative data provided by the Maryland State Department of Education.

Figure 1. Distribution of average confidence interval widths for schools' K–3 math growth estimates, by whether all students or a subset of students take the Kindergarten Readiness Assessment (KRA)

Number of schools



Note: See appendix A for details on how confidence intervals were calculated. This figure shows results using the same sampling percentages that the Maryland State Department of Education (MSDE) used in the 2016/17 school year (shown in table 2).
Source: Administrative data provided by MSDE.

Limitations

The findings and conclusions in this study reflect only the students who had a KRA score and a grade 3 PARCC score, both of which are required to calculate SGP. Students included in the analysis had higher KRA scores and higher PARCC scores than the students in the original study sample who were excluded from the analysis (tables B.1 through B.4 in appendix B). If growth could be calculated for students missing one or both of these scores, SGP estimates might change for the schools in which these students were enrolled. The findings and conclusions may also differ under different growth models, such as a value-added models (VAM), which are used by other states to calculate growth.

We could not measure year-to-year stability (that is, reliability) for schools' K–3 growth estimates because the 2017/18 grade 3 cohort is the first cohort to have completed the KRA. Examining year-to-year reliability can boost schools' confidence in the measure by demonstrating that estimates are not driven by random year-to-year fluctuation in the data (Goldschmidt, Choi, & Beaudoin, 2012).

The growth estimates calculated in this study might be biased because student movement from one school to another within school years could not be observed in the data. Because the data only listed a single school for each student in each school year, the growth estimates account for student movement between school years, but not within school years. Appendix B provides information on the extent to which students moved within school years. Collecting and maintaining data on student movement within a school year, such that there is attendance data on all schools that students attend during each school year, would enable states to account for this type of movement when estimating schools' K–3 growth.

Finally, states may decide to alter or adopt new statewide assessments after establishing a K–3 growth measure. For example, beginning in 2020, Maryland will replace the PARCC assessments with a new statewide assessment. The findings of this study may differ from the findings of a similar study using future student growth data using the newer assessment(s). The methods used in this study also provides states with a blueprint for examining how the validity and precision of their measure changes with the introduction of a new or revised assessment.

Implications

The study's findings suggest that K–3 estimates are likely less valid than later grade estimates in Maryland. This study provides a starting point to understanding a way to measure schools' contributions to student academic growth in the early grades in Maryland, and provides a model to other states interested in constructing similar measures. To effectively guide policy and resources, states need information about how their elementary schools are supporting student learning in all grades. This study breaks new ground by estimating K–3 growth using two different assessments. Such measures can provide states with critical information about elementary schools' performance, but states need to consider the validity and precision of such estimates before they are used for school accountability.

The study's findings suggest that schools' K–3 growth estimates are likely less valid than schools' grades 3–4 growth estimates, because correlations between KRA and grade 3 PARCC scores were smaller than correlations between PARCC scores in later grades, but a comparison to other studies suggests that the KRA is predicting grade 3 achievement reasonably well, relative to other kindergarten assessments. Findings from other studies suggest that assessments of young children may be prone to lower correlations with later assessments. For example, a range of kindergarten readiness measures had weak to moderate correlations with grade 3 achievement scores, ranging from 0.10 to 0.61 for grade 3 reading and 0.12 to 0.68 for grade 3 math, in two large United States studies (Duncan et al, 2007). The highest correlations from these studies are from a measure of math skills at kindergarten entry that was designed to track development over time in the 1998 Early Childhood Longitudinal Study (ECLS-K), and these correlations are only slightly higher than the correlations between KRA and grade 3 PARCC scores. If K–3 estimates are prone to be less valid than later grade estimates, states could acknowledge these differences by assigning less weight to the K–3 growth measure in their accountability framework, relative to growth measures in later grades. States could also consider publicly reporting the K–3 growth measure, but not using it in their formal accountability system.

The study's findings also suggest schools' K–3 growth estimates are much less precise for smaller schools than for larger schools in Maryland. States might want to ensure that resource decisions based on schools' K–3 growth estimates take into account the confidence interval around those estimates. Approaches have been used to improve precision for samples such as small schools. For example, a method called empirical Bayes shrinkage can adjust imprecise growth estimates to be closer to the average estimate across all schools, reducing the chance that schools with less precise estimates (such as small schools) incorrectly receive particular rewards or consequences because of having an extreme but imprecise growth estimate (Herrmann, Walsh, & Isenberg, 2016). Alternatively, or in addition, states could publish schools' growth estimates only after multiple years of data are available. The precision of schools' growth estimates would increase by using data on multiple cohorts of students in each school. Properly incorporating uncertainty into education decision making is a challenge that is the subject of ongoing research (for an example, see Resch, 2017).

Finally, the study's findings suggest that administering the KRA to a subset of students greatly reduces the precision of schools' K–3 growth estimates. Recognizing that assessments like the KRA might be burdensome because they are administered to students one on one by teachers, states could explore strategies that would provide sufficient resources to districts to support a requirement that all students take the assessment.

References

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51. Retrieved from ERIC database, https://eric.ed.gov/?q=Norm-+and+criterion-referenced+student+growth&id=EJ866087.

Betebenner, D., VanIwaarden, A., Domingue B., & Shang, Y. (2019). *SGP: An R package for the calculation and visualization of student growth percentiles and percentile growth trajectories* (R package version 1.9–0.0). Retrieved from http://centerforassessment.github.io/SGP/.

Carroll, R. J., Maca, J. D., & Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika, 86,* 541–554.

Castellano, K. E., & Ho, A. D. (2015). Practical differences among aggregate-level conditional status metrics: From median student growth percentiles to value-added models. *Journal of Educational and Behavioral Statistics, 40*(1), 35–68. Retrieved from ERIC database, https://eric.ed.gov/?q=Practical+differences+among+aggregate-level+conditional+status+metrics&id=EJ1049833.

Castellano, K. E., & McCaffrey, D. F. (2017). The accuracy of aggregate student growth percentiles as indicators of educator performance. *Educational Measurement: Issues and Practice, 36*(1), 14–27. Retrieved from ERIC database, https://eric.ed.gov/?q=The+accuracy+of+aggregate+student+growth+percentiles+as+indicators+of+educator+performance&id=EJ1135072.

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association, 89,* 1314–1328.

Dancey, C.P. and Reidy, J. (2007). Statistics without Maths for Psychology. Pearson Education.

De Boor, C. (2001). *A practical guide to splines*. New York: Springer.

Duncan, G.J., Dowsett, C.J., Claessens, A., Magnuson, K., Huston, A.C., Klebanov, P., … Japel, C. (2007). School readiness and later achievement. Developmental Psychology, 43(6), 1428–1446. Retrieved from ERIC database, https://eric.ed.gov/?q=School+readiness+and+later+achievement&id=EJ779938.

Goldschmidt, P., Choi, K., & Beaudoin, J.P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance*. Washington, DC: Council of Chief State School Officers. Retrieved from ERIC database, https://eric.ed.gov/?q=Growth+model+comparison+study%3a+Practical+implications+of+alternative+models+for+evaluating+school+performance&id=ED542761.

Hardle, W., Muller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and semiparametric models*. New York: Springer.

Herrmann, M., Walsh, E. & Isenberg, E. (2016). Shrinkage of value-added estimates and characteristics of students with hard-to-predict achievement levels. *Statistics and Public Policy, 3*(1), 1–10.

Isenberg, E., & Walsh, E. (2014). *Measuring teacher value added in DC, 2013–2014 school year*. Washington, DC: Mathematica Policy Research.

Jacob, B. A., Lefgren, L., & Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of Human Resources, 45*(4), 915–943. Retrieved from ERIC database, https://eric.ed.gov/?q=The+persistence+of+teacher-induced+learning&id=EJ900945.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning.* New York: Springer.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (NBER Working Paper No. w14607). Cambridge, MA: National Bureau of Economic Research. Retrieved from ERIC database, https://eric.ed.gov/?q=Estimating+teacher+impacts+on+student+achievement&id=ED503840.

Koenker, R. (2005). *Quantile regression*. Cambridge, United Kingdom: Cambridge University Press.

Maryland State Department of Education. (2018). *Maryland Every Student Succeeds Act (ESSA) consolidated state plan final.* Retrieved from http://marylandpublicschools.org/about/Pages/DAPI/ESSA/index.aspx.

Maryland State Department of Education, & Ready at Five. (2017). *Readiness Matters: Informing the future, the 2016–2017 Kindergarten Readiness Assessment technical report.* Retrieved from ERIC database, https://eric.ed.gov/?q=Readiness+Matters%3a+Informing+the+future&id=ED589984.

Maryland State Department of Education, & Ready at Five. (2018). *Readiness Matters: Equity Matters, the 2017–2018 Kindergarten Readiness Assessment report.* Retrieved from ERIC database, https://eric.ed.gov/?q=Maryland+State+Department+of+Education&id=ED589988.

Maryland State Department of Education, & Ready at Five. (2019). *Readiness Matters, the 2018–2019 Kindergarten Readiness Assessment report.* Retrieved from https://earlychildhood.marylandpublicschools.org/system/files/filedepot/4/2018-19_rm_book.pdf.

Maryland State Department of Education. (n.d.). *Student report: Kindergarten Readiness Assessment.* Retrieved from http://www.marylandpublicschools.org/about/Documents/DAAIT/KRA/KRAISR.pdf.

McCaffrey, D. F., Castellano, K. E., & Lockwood, J. R. (2015). The impact of measurement error on the accuracy of individual and aggregate SGP. *Educational Measurement: Issues and Practice, 34*(1), 15–21. Retrieved from ERIC database, https://eric.ed.gov/?q=The+impact+of+measurement+error+on+the+accuracy+of+individual+and+aggregate+SGP&id=EJ1054132.

O'Keefe, B. (2017). *The state of early learning in ESSA: Plans and opportunities for implementation*. New Brunswick, NJ: Center on Enhancing Early Learning Outcomes, & the Council of Chief State School Officers. Retrieved from ERIC database, https://eric.ed.gov/?q=The+state+of+early+learning+in+ESSA%3a+Plans+and+opportunities+for+implementation&id=ED582919.

Partnership for Assessment of Readiness for College and Careers. (n.d.) *English language arts/literacy and mathematics sample score reports.* Retrieved from https://parcc-assessment.org/content/uploads/released_materials/06/PARCC%20Sample%20ISR%2063016.pdf.

Pearson. (2019). *Final technical report for 2018 administration*. Retrieved from: https://parcc-assessment.org/wp-content/uploads/2019/05/PARCC-2018-Technical-Report_Final_02282019_FORWEB.pdf.

R Development Core Team. (2019). *R: A language and environment for statistical computing* (3-900051-07-0). Vienna, Austria: R Foundation for Statistical Computing.

Regional Educational Laboratory Northwest. (2017). *50 state scan of kindergarten readiness definitions and assessments*. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/northwest/pdf/50-state-scan-kindergarten-readiness.xlsx.

Resch, A. (2017, March 31). Moving beyond p-values to help school districts make smarter decisions [Blog post]. Retrieved from https://www.brookings.edu/blog/brown-center-chalkboard/2017/03/31/moving-beyond-p-values-to-help-school-districts-make-smarter-decisions/.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics, 125*(1), 175–214. Retrieved from https://eric.ed.gov/?q=Teacher+quality+in+educational+production%3a+Tracking%2c+decay%2c+and+student+achievement&id=ED503116.

Shang, Y., VanIwaarden, A., & Betebenner, D. W. (2015). Covariate measurement error correction for student growth percentiles using the SIMEX method. *Educational Measurement: Issues and Practice, 34*(1), 4–14. Retrieved from ERIC database, https://eric.ed.gov/?q=Betebenner&ff1=autBetebenner%2c+Damian+W.&id=EJ1054129.

Walsh, E., & Isenberg, E. (2015). How does value added compare to student growth percentiles? *Statistics and Public Policy, 2*(1).

Walsh, E., Liu, A. Y., & Dotter, D. (2015). *Measuring teacher and school value added in Oklahoma, 2013–2014 school year*. Washington, DC: Mathematica Policy Research.

WestEd. (2014). *Kindergarten Readiness Assessment technical report fall 2014*. Retrieved from https://education.ohio.gov/getattachment/Topics/Early-Learning/Kindergarten/Ohios-Kindergarten-Readiness-Assessment/Kindergarten-Readiness-Assessment-for-Data-Manager/KRA_Technical_Report_2014_Final.pdf.aspx.

December 2019

This report was prepared under Contract ED-IES-17-C-0006 by Regional Educational Laboratory Mid-Atlantic administered by Mathematica. The report has been peer reviewed and approved for publication by the U.S. Department of Education’s Institute of Education Sciences. The content of the publication does not necessarily reflect the views or policies of the Institute of Education Sciences or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

This REL report is in the public domain. While permission to reprint this publication is not necessary, it should be cited as:

Dragoset, L, Baxter, C., Dotter, D., & Walsh, E. (2019). *Measuring school performance for early elementary grades in Maryland.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

Appendix A. Methods

A detailed description of the data, sample, and methods used in this study is provided below.

Data

A full list of the data files from the Maryland State Department of Education (MSDE) is provided in table A.1. Students, schools, and districts were uniquely identified across these files using student, school, and district identification (ID) numbers that MSDE provided. No personal identifiers such as student names were included in the data.

Kindergarten Readiness Assessment (KRA) data. For each kindergartner administered the KRA in 2014, the data included an overall KRA scaled score; a readiness level based on this overall score (demonstrating, approaching, and emerging); and a scaled score for each of the assessment's four domains: language and literacy, mathematics, social foundations, and physical well-being and motor development. The data file also indicated the school and district where the student took the KRA. (These data excluded scores for students at the Maryland School for the Blind and the Maryland School for the Deaf.) See table A.2 for a description of the KRA assessment.

Partnership for Assessment of Readiness for College and Careers (PARCC). For each student, the dataset included scaled scores for the reading and math subject tests that are administered as part of the spring PARCC in grades 3 through 6, and a proficiency level for each subject based on these scores (ranging from 1 to 5; see table A.2 for a description of the PARCC assessment). The data file also indicated the school and district where the student took the spring PARCC.

Attendance data. For each student year, the data included the grade and school in which the student was enrolled as of the last day of the school year, the date the student enrolled in the school that year, and the number of days in attendance and absent.

Discipline data and student background characteristics. The dataset included student discipline data and background characteristics that were used to conduct supplemental analyses that examined the characteristics of students included versus excluded from the analyses.

Table A.1. Data provided by the Maryland State Department of Education

Data source	How data are used
KRA scores for 2014/15 school year	Kindergarten scores for the K–3 SGP
PARCC scores for 2014/15 through 2017/18 school years	Grade 3 scores for the K–3 SGP; Comparisons of K–3 SGPs and SGPs for grades 4 through 6
Attendance data for 2014/15 through 2017/18 school years	Link students to schools and districts and provide most up-to-date data on student background characteristics used in the analysis of students included versus excluded from the analyses
Discipline data for 2014/15 through 2017/18 school years	Provide infraction data used in the analysis of students included versus excluded from the analyses

KRA is Kindergarten Readiness Assessment. PARCC is Partnership for Assessment of Readiness for College and Careers. SGP is student growth percentile.
Note: All data files contain student background characteristics that were used from the data sources in the following order: (1) the attendance files; (2) data in the assessment files (PARCC or KRA); and (3) data in the discipline files. These characteristics were used in the analysis of students included versus excluded from the sample.

Table A.2. Snapshot of the KRA and PARCC assessments

Assessment	When it is administered	What it measures	How it is scored	How it is scaled
KRA	Kindergarten, fall	Kindergarten readiness across four domains: language and literacy, mathematics, social foundations, and physical well-being and motor development	Domain subscores range from 202 to 298 (and to 293 for physical well-being and motor development) Overall scaled score ranges from 202 to 298	Overall scores are categorized into 3 performance levels: *Demonstrating readiness*: A child demonstrates foundational skills and behaviors that prepare him/her for curriculum based on kindergarten standards. *Approaching readiness*: A child demonstrates some foundational skills and behaviors that prepare him/her for curriculum based on kindergarten standards. *Emerging readiness*: A child demonstrates minimal foundational skills and behaviors that prepare him/her for curriculum based on kindergarten standards.
PARCC	Grades 3—8, spring[a]	Mastery of grade-specific standards for English language arts/literacy (reading) and math	Overall scaled score for each subject (reading and math) range from 650 to 850	Subject scaled scores are categorized into 5 performance levels: *Level 5*: Exceeded expectations *Level 4*: Met expectations *Level 3*: Approached expectations *Level 2*: Partially met expectations *Level 1*: Did not yet meet expectations

[a] End-of-course assessments in English, Algebra I, Algebra II, and Geometry are administered in high school, and excluded here because they are not used in MSDE's SGP models.

KRA is Kindergarten Readiness Assessment. PARCC is Partnership for Assessment of Readiness for College and Careers. SGP is student growth percentile.

Note: Reliability estimates for the KRA and PARCC are provided in technical reports. The raw score reliability coefficients for the assessments used in the model are as follows: 0.95 for the 2014/15 KRA overall score, 0.90 for the grade 3 PARCC reading assessment (0.88 for paper-based), and 0.94 for the grade 3 PARCC math (0.93 for paper-based) (Pearson, 2019; WestEd, 2014).

Source: Maryland State Department of Education (n.d.) and Partnership for Assessment of Readiness for College and Careers (n.d.)

Sample

In collaboration with MSDE, business rules were established to clean the data and define the samples of students eligible for inclusion in the SGP models. These rules were applied to the data to identify the sample of students eligible for inclusion in the K–3 SGP model, and the sample of grades 3–6 assessment scores that were used to evaluate the validity and precision of the K–3 estimates (research question 2). Each of the samples is described below.

K–3 SGP sample

K–3 school sample. Schools must have had at least 10 eligible students for growth estimates to be reported, per MSDE's Every Student Succeeds Act plan. (Definitions for eligible students are defined below.) Among the 977 schools observed serving K–3 students, 905 schools had at least 10 eligible students. Schools with fewer than 10 eligible students are excluded when reporting the distribution and summary statistics for the K–3 growth estimates in figure A.1 and table B.8, respectively, but included in analyses for research questions 2 through 4, so that the statistical properties of these estimates can be fully explored.

Students. A total of 80,860 students were observed with a 2014/15 KRA score record or a 2017/18 grade 3 PARCC/MSAA score record, or were enrolled in kindergarten in 2014/15, grade 1 in 2015/16, grade 2 in 2016/17, or grade 3 in 2017/18 in the student attendance data, after preliminary data cleaning.

Among these students, to be eligible for inclusion in the model, students needed a 2014/15 KRA score and at least one 2017/18 grade 3 PARCC score in math or reading. Because a 2014/15 KRA score and a 2017/18 PARCC score is required to calculate growth, the model necessarily excludes students who were in kindergarten in 2014/15 and then skipped or repeated a grade before grade 3, and students who joined the MSDE school system after their cohort's kindergarten assessment window and/or left the system before their cohort's grade 3 assessment window. Students who took only the alternate assessment (that is, the Multi-State Alternate Assessment, or MSAA) were also excluded.

The number of students who were excluded from the analyses based on these rules is shown in Table A.3. Among all 80,860 students observed, roughly two-thirds were eligible for the model: 54,393 students for the math model and 54,397 for the reading model. Among the original cohort of students who completed the KRA in fall 2014/15 (that is, the group of students for which growth could have been calculated, if the student also completed the grade 3 PARCC), about 86 percent of the cohort was included in the model. About 14 percent were excluded from the model because they lacked a grade 3 assessment score (for example, because they left the Maryland public schools or experienced non-traditional grade progression after taking the KRA). A very small percentage (0.37 percent) were excluded because they took the alternate assessment instead of the PARCC.

Table A.3. Number of students included in and excluded from the K–3 student growth percentile analyses, by subject

	Math		Reading	
Student description	Number	Percentage	Number	Percentage
Students in kindergarten in 2014/15, grade 1 in 2015/16, grade 2 in 2016/17, or grade 3 in 2017/18	80,860	100.00	80,860	100.00
Students with no 2014/15 KRA score[a]	17,342	21.45	17,342	21.45
Students with no 2017/18 grade 3 assessment score (PARCC or alternate)[b]	8,890	10.99	8,886	10.99
Students who took the alternate assessment, rather than the PARCC	235	0.29	235	0.29
Total students excluded	**26,467**	**32.73**	**26,463**	**32.73**
Total students included	**54,393**	**67.27**	**54,397**	**67.27**

KRA is Kindergarten Readiness Assessment. PARCC is Partnership for Assessment of Readiness for College and Careers.

Note: By construction, rows 2–4 sum to row 5; in other words, a student who was excluded from the analysis for multiple reasons only appears in one of rows 2–4, in the first row that applies to that student.

a. These are students who either (1) did not appear in the 2014/15 KRA file (for example, because they entered the Maryland public school system after the 2014/15 KRA was administered), or (2) did appear in the 2014/15 KRA file, but had a missing KRA score because they did not complete the assessment.

b. These are students who (1) have no 2017/18 assessment score (for example, because they left the Maryland public school system before the 2017/18 assessment was administered), (2) have a 2017/18 score but the test code field associated with that score was not grade 3, or (3) had a non-normal grade progression at any point from 2014/15 through 2017/18.

Source: Administrative data provided by the Maryland State Department of Education.

Grades 3–6 secondary analysis sample

Analyses on the validity and precision of the K–3 SGP estimates (research question 2) and supplemental analyses to inform the construction of the SGP model (as described in the Methods section below) were conducted using a sample of 359,619 students with scores on any of the 2014/15 to 2016/17 grade 3, 2014/15 to 2017/18 grade 4 and 5, or 2015/16 to 2017/18 grade 6 PARCC. Among this sample, students were included in each of the analyses if they had a PARCC score in each of the grades being examined. For example, when examining the validity of the K–3 estimates under research question 2, students needed a grade 3 PARCC score and either a grade 4 PARCC score or grade 6 PARCC score to be included in the pairwise correlations for these analyses. In total, scores from five cohorts were examined, as depicted below:

PARCC scores examined:	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Cohort 5
Grade 3			2014/15	2015/16	2016/17
Grade 4		2014/15[a]	2015/16	2016/17	2017/18
Grade 5	2014/15[a]	2015/16[a]	2016/17[a]	2017/18[a]	
Grade 6	2015/16[a]	2016/17[a]	2017/18		

[a] Used in supplemental analyses only

Methods

The following section provides a detailed description of the methods used to calculate schools' growth estimates and to explore each of the study's four research questions.

Calculating SGP estimates and school-level growth estimates

An SGP represents the percentile rank of a student's current score relative to other students who had the same (or a similar) prior score as that student (Betebenner, 2009). For example, a student-level SGP of 75 indicates that the student scored higher than three-quarters of the students who had similar academic achievement at baseline.

The study calculated SGP estimates as described in Betebenner (2009) and used the simulation extrapolation (SIMEX) method as described in Shang, VanIwaarden, & Betebenner (2015) to adjust for measurement error in the baseline assessment score (that is, the extent to which assessment scores do not reflect students' actual ability). The study then transformed the SIMEX-adjusted SGPs into percentile-ranked SIMEX-adjusted SGPs as described in McCaffrey, Castellano, & Lockwood (2015). The percentile-ranked SIMEX-adjusted SGPs were aggregated to the school level by calculating the mean SGP among the students who attended the school. Students who transferred schools between the start of kindergarten and the end of grade 3 contribute to the average growth estimate for each school they attended, weighted proportional to the amount of time the student was enrolled in the school. The model differs from MSDE's SGP model for grades 4–8 in its use of mean SGPs, proportional weights, and an adjustment for measurement error.

The following section details: (1) how the study calculated SGPs, (2) how and why the study implemented the SIMEX method, (3) why the study transformed the results into percentile-ranked SIMEX-adjusted SGPs, (4) how and why mean SGPs with weights proportional to student enrollment were calculated, and (5) results confirming that the SGP estimation methods worked as expected. Specifically, the study confirmed that: the unadjusted SGPs were uncorrelated with students' KRA scores (their baseline assessment score); the adjusted percentile-ranked SIMEX-adjusted SGPs were negatively correlated with students' KRA scores; and schools K–3 growth estimates are dependent upon which students attended each school (suggesting the estimates reflect the average growth of students in the school, as expected).

Calculating SGP estimates. Estimating an SGP involves two steps: (1) estimating quantile regressions of students' current scores on their prior scores, and (2) calculating students' SGP estimates using the results from those regressions. In the first step, the study estimated quantile regressions (as described in Koenker, 2005) of students' current scores on their prior scores.[1] Let Q_i^τ denote the τ th quantile of the current score, given student i's value of the prior score, where i runs from 1 to the number of students (n). For example, for $\tau = 0.50$ (the median), Q_i^τ denotes the median current score among all students with a particular prior score. In the second step, the study compared each student's current score to the τ distinct values of Q_i^τ, found the two consecutive values of Q_i^τ that surround the student's current score, took the midpoint of those two τ values, and multiplied that by 100. For example, if student i's current score lay between $Q_i^{0.935}$ and $Q_i^{0.945}$, that student's estimated SGP was 94.

The study estimated SGPs using the SGP package (Betebenner, VanIwaarden, Domingue, & Shang, 2019) in the statistical software program R (R Development Core Team, 2019). In that package, Q_i^τ is, by default, estimated for $\tau = 0.005, 0.015, \ldots, 0.995$. The package estimates Q_i^τ with nonparametric quantile regression, specifically,

[1] Whereas a linear regression of students' current scores on their prior scores estimates the average change in the current score associated with a one unit increase in the prior score, a quantile regression of students' current scores on their prior scores estimates the change in a specified quantile of the current score associated with a one unit increase in the prior score. For example, a median regression (the median is the 50th percentile) of students' current scores on their prior scores estimates the change in the median current score associated with a one unit increase in the prior score.

quantile regression with cubic B-splines. Unlike parametric models that estimate a single regression for the entire dataset, a regression B-spline divides the dataset into multiple bins of students with similar prior scores and estimates a separate regression within each bin (De Boor, 2001; Hardle, Muller, Sperlich, & Werwatz, 2004). In the SGP R package, the type of regression that is estimated within each bin is a cubic polynomial (meaning that current scores are regressed on prior scores, prior scores squared, and prior scores cubed). The points that divide the data into bins are called knots. The regression functions on either side of each knot are constrained to have the same value and slope at the knot where they meet, so that overall regression function is smooth at the knots. In the SGP R package, by default, the knots are defined as the 20th, 40th, 60th, and 80th percentiles of the prior score.

Accounting for measurement error in the KRA score using the SIMEX method. Measurement error in students' prior scores results in biased SGP estimates for students and biased mean SGP estimates for schools (Castellano & McCaffrey, 2017; Shang et al., 2015). In particular, SGP and mean SGP estimates tend to be underestimated for students or schools with low prior achievement and overestimated for students or schools with high prior achievement. Based on consultations with MSDE, the K–3 model adjusts for measurement error in students' prior scores using the SIMEX method.

The SIMEX method is a measurement error correction technique, originally proposed by Cook and Stefanski (1994), that reduces this bias in SGP and mean SGP estimates (Castellano & McCaffrey, 2017; Shang et al., 2015). The technique first uses simulation to measure the consequences for the quantile score estimates, Q_i^τ, of adding incrementally more measurement error to the prior assessment scores. Then, the technique uses the relationship between the amount of measurement error and the quantile score estimates to extrapolate how the quantile estimates would appear had there been no measurement error in the prior assessment scores. Executing the method consists of two steps: simulation and extrapolation.

In the simulation step, an increasing amount of extra measurement error is added to students' KRA scores, and the quantile scores are calculated for each level of error. This study used four levels of error, described in more detail below. For each of those four levels of measurement error, the following process is repeated. First, a distribution, or set of likely values, for the measurement error is chosen to correspond to the level of measurement error, so that larger errors are more likely when the amount of error is higher. Second, a particular random amount of error from that distribution is chosen for each student and added to that student's KRA score repeatedly 100 times, producing 100 new datasets of students' prior assessment scores. Third, using each of the 100 datasets, quantile scores Q_i^τ are then calculated using the standard method described above. Lastly, the quantile scores are averaged across the 100 iterations, resulting in a single set of quantile scores for the level of measurement error. The study assumed measurement error is distributed according to a normal distribution with a mean of zero and used a value of 10.14 for σ^2, the variance of the measurement error in the observed prior assessment score data. This value is the variance of measurement error for the 2014 overall KRA score based on the reported standard error of measurement and reliability coefficient (WestEd, 2014). The variance of the measurement error added to students' scores for each level of simulated measurement error is equal to $\lambda\sigma^2$, where λ is a number chosen between 0 and 2. For each level of simulated measurement error, the total measurement error (original measurement error plus added measurement error) has a variance equal to $(1+\lambda)\sigma^2$). This study used four values of λ, set to 0.5, 1, 1.5, and 2, following Carroll, Maca, and Ruppert (1999).

In the extrapolation step, the relationship between the averaged quantile scores for each level of measurement error and the value of λ corresponding to that level is measured and used to understand how the quantiles might appear, had no measurement error been present. The relationship is measured using an ordinary least squares regression, and the coefficient from that regression is used to calculate the predicted value of each quantile score at $\lambda = -1$, which is the level of λ that corresponds to a measurement error variance of zero (because $(1+\lambda)\sigma^2$

is equal to 0 when $\lambda = -1$). These predicted values are the SIMEX estimates of the quantile scores, denoted $Q^{\tau}_{i,\, SIMEX}$. The SGP estimate derived from $Q^{\tau}_{i,\, SIMEX}$ is denoted as SGPSIMEX.

Transforming the results into percentile-ranked SIMEX-adjusted SGPs. Following McCaffrey et al. (2015), the study makes one final adjustment to the SIMEX-adjusted SGP estimates, to improve the distributional properties of those estimates. In particular, the study takes the percentile ranks of the SIMEX-adjusted SGP estimates, which is an ordinal transformation that preserves the ordered ranking of students' SGP estimates. McCaffrey et al. demonstrate that SIMEX-adjusted SGPs have a U-shaped distribution, meaning that larger percentages of students have SGP values near 0 and 100 than near 50. In contrast, non-SIMEX-adjusted SGPs, as well as SGPs based on an assessment without any measurement error, have the appealing property of being uniformly distributed (meaning that the percentage of students with a particular SGP value is the same for every value). McCaffrey et al. show that taking the percentile ranks of SIMEX-adjusted SGPs results in SGP estimates that are uniformly distributed and that retain the desirable property of SIMEX-adjusted SGPs (that is, they reduce the bias caused by measurement error in the prior assessment).

Calculating schools' mean SGP estimates.

School-level growth estimates are constructed as weighted averages of their assigned students' SGPs. The study calculated mean SGPs, rather than median SGPs, for two reasons: (1) our analyses showed mean SGPs are more precise than median SGPs (table A.7 compares the average confidence interval widths for the median SGPs and mean SGPs estimates) and (2) prior research using simulations has shown that mean SGPs are less biased than median SGPs (Castellano & Ho, 2015; Castellano & McCaffrey, 2017). The model accounts for student movement across schools because the opportunity for movement over the K–3 period, which spans four grades, is significantly higher than the amount expected for the grade 4–8 SGPs, which each span only onegrade.

To account for student movement between schools, the model assigns weights proportional to the amount of time that the student was enrolled at each school over the K–3 period. Proportional weighting comprises three steps: (1) identifying each school that the student attended over the period, (2) determining the proportion of the K–3 period that the student was enrolled at each school, and (3) constructing the final weights as described below:

1. **Identifying schools.** Ideally, every school that the student attended between administration of the KRA and the grade 3 PARCC would receive credit for a portion of the student's estimated academic growth over this period (that is, the student's SGP estimate). However, a complete attendance record for each student was not part of the data available for this study. Rather, students' school enrollment in the dataset is observed at six points in time: (1) at KRA administration, (2) on the last day of kindergarten, (3) on the last day of grade 1, (4) on the last day of grade 2, (5) at PARCC administration, and (6) on the last day of grade 3.

   The student attendance data identify the school where the student was enrolled as of the last day of the school year in kindergarten through grade 3 (hereafter referred to as the student's end-of-year school), the date the student enrolled in the school that year, and the number of days the student was in attendance and absent. The assessment data identify the school where the student was enrolled when he or she completed the KRA and grade 3 PARCC (hereafter referred to as the student's assessment school). Multiple transfers between these two time points and attendance at other schools before the assessments in a given year are not observed.

   To calculate schools' growth estimates, students are assigned to the single school in each school year for which the number of days they were in attendance or absent is available: their end-of-year school. The one exception to this rule is that students are assigned to their grade 3 assessment school for grade 3, even when that school differs from their end-of-year school that year. In the latter case, students transferred to their end-of-year school after taking the grade 3 assessment, and the end-of-year school did not contribute to the academic growth measured by this model. Such students are therefore not assigned to their end-of-year school for the grade 3 school year to prevent them from contributing to a growth estimate for a school they did not attend between the KRA and grade 3 PARCC. This includes instances where the student has a grade 3 assessment score but is not observed in the grade 3 attendance data file, indicating the student exited the MSDE school system after taking the PARCC.

2. **Calculating the percentage of time enrolled.** Students' percentage of days enrolled in their assigned school for each year is defined as the number of days they were enrolled in the school that year out of the total number of days the school was in session the same year.

   The number of days enrolled is defined as the sum of the total days attended and the total days absent, as indicated for student's end-of-year school on the attendance data provided by MSDE. The total number of days the school was in session is defined as 180 days, which is the number of instructional days for the MSDE school year. When attendance records showed a student as being enrolled for more than 180 days, the student is assumed to have been enrolled for 100 percent of the school year. (This rule impacts 24 percent of all student-year records in the K–3 sample, but the differences are small: most are observed as enrolled for 181 days and the maximum is 188 days.)

   Because the student attendance data only reports days attended and days absent for the student's end-of-year school, data on students' total days enrolled in the assessment school are not available when the school differs from their end-of-year school. In these cases (less than 1 percent of students in the K–3 sample), the percentage of days enrolled in the assessment school is therefore estimated and assumes the student was enrolled in his or her assessment school for the portion of the year that the student was not enrolled in his or her end-of-year school. That is, when the two schools differ, the percentage of days enrolled at the grade 3 assessment school is calculated as 100 percent minus the percentage of days enrolled in the grade 3 end-of-year school. Eighteen students in the K–3 sample have a value of zero for the percentage of days enrolled at their grade 3 school (meaning that these students do not contribute to schools' growth estimates for their grade 3 year, though they do contribute for other years), because their grade 3 end-of-year school differs from their grade 3 assessment school, and the attendance data provided by MSDE indicate these students were enrolled at their end-of-year school for the entire year.

3. **Constructing proportional weights.** The estimated SGP of each of the school's students is weighted according to the amount of time the student was enrolled in the school over the four-year period. Specifically, each estimated SGP is apportioned to the student's assigned school in each year with a weight equal to the proportion of time the student was enrolled in the school that year multiplied by 0.25, to reflect that growth is estimated over four years. Students' growth estimates could be more precisely apportioned across schools by using more comprehensive student enrollment data, including whether a student attended part-day or full-day kindergarten, and with detailed information on the timing of the PARCC in each school.

   For example, a student who enrolled in kindergarten at the start of the year and attended the same school through grade 3, without ever transferring schools, would be observed with 100 percent days enrolled at the same school in each of the four years, as indicated below. Therefore, this student's estimated SGP would be entirely attributed to school ID 100.

Grade	School ID	Percent of days enrolled	Percent of four-year period	Weight
K	100	100	100/4=25	1 * 0.25 = 0.25
1	100	100	100/4=25	1 * 0.25 = 0.25
2	100	100	100/4=25	1 * 0.25 = 0.25
3	100	100	100/4=25	1 * 0.25 = 0.25

Next, take an example of a student who enrolled in kindergarten at the start of year and attended the same school through grade 2, but then changed schools *between* grades 2 and 3. This student would be also observed with 100 percent days enrolled in each of the four years, but his or her school ID would change in grade 3. Therefore, 75 percent of the student's estimated SGP would be attributed to school 200, which taught the student for three full years, and 25 percent would be attributed to school 201, which taught the student one full year.

Grade	School ID	Percent of days enrolled	Percent of four-year period	Weight
K	200	100	100/4=25	1 * 0.25 = 0.25
1	200	100	100/4=25	1 * 0.25 = 0.25
2	200	100	100/4=25	1 * 0.25 = 0.25
3	201	100	100/4=25	1 * 0.25 = 0.25

Next, take an example of a student who enrolled in kindergarten at the start of the year and then changed schools (from school 300 to school 301) in the middle of grade 1. This student would be observed with 100 percent days enrolled in kindergarten for school 300, 100 percent days enrolled in grades 2 and 3 for school 301, but only 50 percent days enrolled in grade 1, because the student enrolled in his or her end-of-year school (school 301) in the middle of the school year. School 301, which taught the student for two and a half years, would receive 62.5 percent of the student's estimated SGP. School 300, which taught the student for one and a half years, should receive the remaining 37.5 percent. However, it would receive only 25 percent, and the remaining 12.5 percent would not be attributed to any school's estimate, because the student's enrollment in school 300 during the first half of grade 1 is unobserved in the data.

Grade	School ID	Percent of days enrolled	Percent of four-year period	Weight
K	300	100	100/4=25	1 * 0.25 = 0.25
1	301	50	100/4=25	0.5 * 0.25 = 0.125
2	301	100	100/4=25	1 * 0.25 = 0.25
3	301	100	100/4=25	1 * 0.25 = 0.25

Finally, take an example of a student who enrolled in kindergarten at the start of the year, left the MSDE school system before the end of the year, but then returned in grade 1 and attended the same school through grade 3. This student would be observed with zero days enrolled in kindergarten (because the student does not have an end-of-year school), and then 100 percent days enrolled in grades 1 through 3. School 401 receives 75 percent of the student's estimated SGP and the remaining 25 percent is unattributed because the school the student attended after taking the KRA is unobserved in the data.

Grade	School ID	Percent of days enrolled	Percent of four-year period	Weight
K		0	0/4=25	0 * 0.25 = 0
1	401	100	100/4=25	1 * 0.25 = 0.25
2	401	100	100/4=25	1 * 0.25 = 0.25
3	401	100	100/4=25	1 * 0.25 = 0.25

These examples illustrate two important points. First, in this study the credit a school receives for a student's estimated growth (that is, the student's estimated SGP) is proportional to the amount of time the student was (observed to be) enrolled in the school over the four-year period, with each of the four years weighted equally. Second, schools only receive credit for *observed* enrollment, and a student's enrollments cannot be completely observed in the data if the student transferred schools during a school year and/or left and then returned to the MSDE school system between the K–3 period.

Before proportional weights were chosen, larger weights for schools attended more recently were explored based on empirical evidence that schools' contributions, and teachers' impacts specifically, dissipate over time (Jacob, Lefgren, & Sims, 2010; Kane & Staiger, 2008; Rothstein, 2010). The study team examined whether the relationship between a student's assessment scores in two grade levels weakened as the time between the grade levels increased, which is a pattern that is consistent with a teacher's impact on a student's performance fading out over time, once the student moves onto a new teacher. This pattern would suggest that learning that occurred more recently, and thus schools that were attended more recently, should receive more weight. Because no statewide assessments in grades 1 and 2 exist that would support this analysis for K–3 directly, the study team used data from the grade 3–6 PARCC. Specifically, students' grade 6 PARCC scores were correlated with their PARCC scores in grade 5, grade 4, and then grade 3, for the cohort of students who were in grade 6 in 2017/18 (cohort 3). To gather confidence that the patterns detected for cohort 3 are roughly representative of other cohorts, the team compared these correlations to the available data for two other cohorts, students who were in grade 6 in 2016/17 (cohort 2) and in 2015/16 (cohort 1). The results of this analysis are presented in table A.4.

Table A.4. Correlations between students' Partnership for Assessment of Readiness for College and Careers (PARCC) scores over time

Correlation between students' PARCC scores in:	Math			Reading		
	Cohort 3	Cohort 2	Cohort 1	Cohort 3	Cohort 2	Cohort 1
Grade 3 and grade 6	0.82			0.77		
Grade 4 and grade 6	0.84	0.85		0.81	0.81	
Grade 5 and grade 6	0.86	0.85	0.85	0.84	0.84	0.82

Note: Data are limited to students in the grade 3–6 analysis sample. Cohorts 3, 2, and 1 refer to students enrolled in grade 6 in 2017/18, 2016/17, and 2015/16, respectively. For each subject (math and reading), the study tested whether (1) the correlation between grade 3 and grade 6 scores (shown in the first row of the table) differed from the correlation between grade 4 and grade 6 scores for the third cohort, (2) the correlation between grade 3 and grade 6 scores differed from the correlation between grade 5 and grade 6 scores for the third cohort, (3) the correlation between grade 4 and 6 scores differed from the correlation between grade 5 and grade 6 scores for the third cohort, (4) the correlations between grades 4 and grade 6 scores differed across cohorts, and (5) the correlations between grade 5 and 6 scores differed across cohorts. All of these tests were significant at the 5 percent significance level, except for the grade 4 and 6 correlations between the second and third cohorts for reading and the grade 5 and 6 correlations between the first and third cohort for math and between the second and third cohort for reading.

Source: PARCC math and reading subject tests for grades 3–6, provided by the Maryland State Department of Education.

In cohort 3, the correlations do decline with each earlier assessment, consistent with the notion that teachers' impacts are dissipating over time between grades 3 and 6, and schools attended more recently should therefore receive a greater weight. (The comparison to cohorts 1 and 2 suggests that these correlations are relatively stable across cohorts). However, the differences are not substantial enough to suggest that proportional weighting might be an inappropriate assumption. Moreover, all of the correlations are within the range of correlations between pre-test and post-test scores used in growth models in other contexts (Isenberg & Walsh, 2014; Walsh, Liu, and Dotter, 2015). Therefore, the straightforward and easily replicated approach of proportional weighting was the most appealing assumption to impose on model weights.

Modifications to the method for weighting students in the analysis are unlikely to have a major effect on the study findings. The study examined the extent to which schools' K–3 growth estimates would change if a simpler weighting method were used. In particular, the study calculated growth estimates in which each student's full weight (of 1) was given to the school in which he or she took the grade 3 assessment. Schools' growth estimates were largely unaffected: the correlation between estimates using the actual weighting approach and estimates using the simpler weighting approach was 0.99, and the SGP estimate for the vast majority of schools changed by less than 5 percentile points.

Confirming that the SGP estimation methods worked as expected. The study conducted three tests to confirm that the SGP estimation methods described above worked as expected. First, the study examined whether students' SGP estimates are uncorrelated with students' KRA scores when no adjustment for measurement error in KRA scores was made. Growth measures are intended to measure growth in a way that gives each student an equal opportunity to achieve high growth, regardless of their score on the initial assessment, and therefore gives schools equal opportunity to obtain a high growth estimate, regardless of whether the school serves students with greater or lesser degrees of kindergarten readiness. So in practice (ignoring measurement error for a moment), a growth estimate that appropriately accounts for the initial assessment score should result in a similar proportion of high-growth students among students with high and low initial assessment scores. The study examined students' K–3 SGP estimates to see if they have this expected pattern.

As expected, the study found that students' SGP estimates are uncorrelated with students' KRA scores when no adjustment for measurement error in KRA scores was made. In particular, table A.5 shows that, regardless of students' KRA scores, they have a roughly equal chance of having low growth, medium growth, and high growth in math (the same was true for reading; results not shown). This is indicated in the table by each cell containing 10 percent (or close to 10 percent) of students. Each row contains students with a certain range of KRA scores; for example, the first row shows students with the lowest KRA scores, ranging from 202 to 254. Each column contains students with a certain range of SGP values; for example, the first column shows students with the lowest growth (SGPs ranging from 1 to 9). That students are spread fairly evenly across cells demonstrates that students are equally likely to have growth of any size, regardless of their baseline assessment score.

Table A.5. Percentage of students who fall within particular ranges of overall Kindergarten Readiness Assessment (KRA) scores and math student growth percentile (SGP) estimates, not accounting for measurement error in KRA scores

Decile/range of overall 2014/15 KRA scale score	Range of math SGP estimates									
	1 to 9	10 to 19	20 to 29	30 to 39	40 to 49	50 to 59	60 to 69	70 to 79	80 to 89	90 to 99
1st/202 to 254	10.4	10.1	9.9	10.3	10.0	9.9	10.4	9.7	9.9	9.3
2nd/255 to 259	11.0	9.8	9.8	9.4	10.9	9.4	9.5	10.4	10.1	9.7
3rd/260 to 263	9.6	10.7	10.9	10.6	8.5	10.7	9.9	10.3	9.8	9.0
4th/264 to 266	11.3	9.4	10.3	9.4	10.3	10.1	10.0	9.4	9.7	10.2
5th/267 to 269	11.0	9.4	9.6	10.4	10.6	9.3	10.8	9.3	10.4	9.1
6th/270 to 272	10.3	9.7	9.5	11.0	10.0	10.6	9.1	10.9	9.9	9.2
7th/273 to 276	10.9	10.3	10.8	8.8	9.9	9.5	10.4	9.7	10.5	9.3
8th/277 to 280	10.4	10.2	9.9	9.8	10.7	9.8	10.2	9.6	10.3	9.1
9th/281 to 286	10.5	10.2	10.0	10.3	9.4	10.0	10.5	10.0	9.5	9.6
10th/287 to 298	10.5	10.0	10.9	10.0	9.8	10.5	9.3	9.6	10.8	8.7

Source: Administrative data provided by the Maryland State Department of Education, 2014/15 to 2017/18.

Second, the study examined whether students' percentile-ranked SIMEX-adjusted SGP estimates (which account for measurement error in KRA scores) are negatively correlated with students' KRA scores. Measurement error causes some students with high ability to incorrectly receive low KRA scores, and vice versa. After adjusting for measurement error, students' K–3 percentile-ranked SIMEX-adjusted SGP estimates should be negatively correlated with students' KRA scores (Shang et al., 2015).

As expected, the study found that students' percentile-ranked SIMEX-adjusted SGP estimates (which account for measurement error in KRA scores) are negatively correlated with students' KRA scores. This is indicated in table A.6 by more than 10 percent of students with the lowest KRA scores having high math SGP estimates, and more than 10 percent of students with the highest KRA scores having low math SGP estimates (the same was true for reading; results not shown).

Table A.6. Percentage of students who fall within particular ranges of overall Kindergarten Readiness Assessment (KRA) scores and math student growth percentile (SGP) estimates, accounting for measurement error in KRA scores

Decile/range of overall 2014/15 KRA scale score	Range of percentile-ranked simulation extrapolation method -adjusted math SGP estimates									
	1 to 9	10 to 19	20 to 29	30 to 39	40 to 49	50 to 59	60 to 69	70 to 79	80 to 89	90 to 99
1st/202 to 254	8.1	9.0	8.0	10.1	9.8	10.0	11.9	10.6	11.0	11.5
2nd/255 to 259	9.1	8.8	9.0	10.2	9.6	9.8	9.3	11.0	11.6	11.6
3rd/260 to 263	8.6	9.9	10.8	9.8	9.3	10.0	10.3	11.0	10.0	10.4
4th/264 to 266	9.9	9.5	9.7	11.3	8.7	10.3	10.4	9.7	10.3	10.2
5th/267 to 269	11.2	9.0	10.3	8.6	12.0	8.6	11.4	9.9	10.0	9.1
6th/270 to 272	10.3	10.0	10.2	11.3	9.7	9.4	10.5	9.5	10.0	9.1
7th/273 to 276	12.4	9.9	11.7	8.6	10.4	9.0	10.0	9.7	9.7	8.8
8th/277 to 280	11.5	11.7	9.5	11.6	10.3	11.5	8.6	8.2	8.7	8.2
9th/281 to 286	12.1	11.8	10.3	9.2	10.4	11.3	8.2	9.6	8.6	8.5
10th/287 to 298	11.4	10.8	9.9	10.9	10.0	9.2	9.5	11.2	8.8	8.2

Source: Administrative data provided by the Maryland State Department of Education, 2014/15 to 2017/18.

Third, the study investigated whether schools' K–3 growth estimates were dependent upon which students attended each school. Observing such a dependency increases one's confidence that schools' growth estimates measure the true average growth of students in the school. The study performed a falsification test in which students were randomly assigned to schools in the sample and schools' K–3 growth estimates were recalculated using these false school assignments. The study then compared the distribution of schools' growth estimates based on the false assignments to the distribution of estimates based on students' actual assignments. Because, under random assignment, students should be assigned to their actual school only by chance, one would expect these distributions to look quite different.

As expected, the study found that schools' K–3 growth estimates are dependent upon which students attended each school, illustrated by the fact that the distribution of schools' growth estimates based on random assignment of students to schools looks quite different from the distribution of schools' growth estimates based on students' actual schools (figure A.1). The distribution based on random school assignments is clustered around 50, whereas the distribution based on actual school assignments is more normally distributed. The average difference between growth estimates based on random school assignments and growth estimates based on actual school assignments was significant at the 5 percent significance level.

Figure A.1. Distribution of schools' K–3 growth estimates, using students' actual schools versus randomly assigned schools

Number of schools

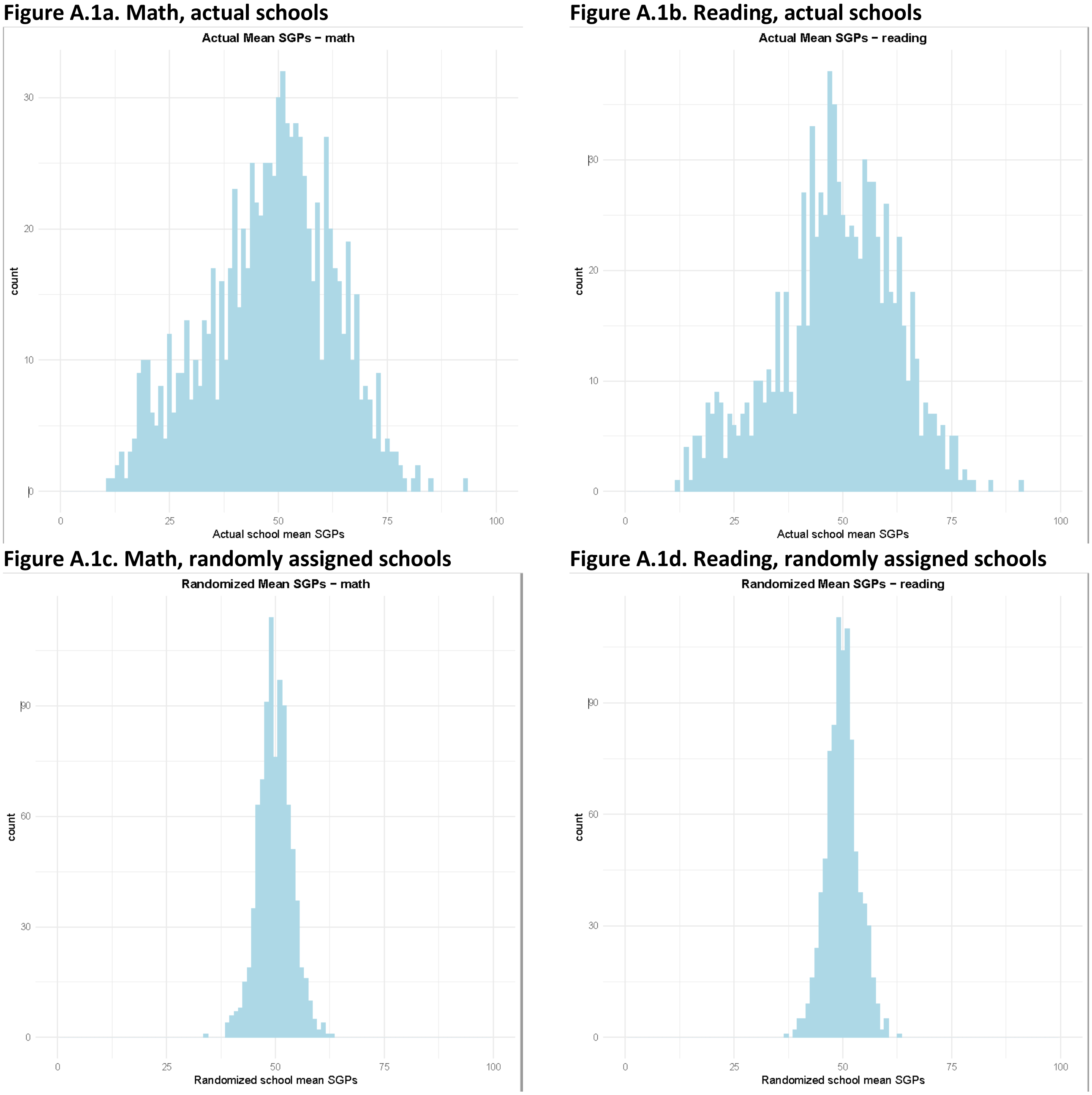


Figure A.1a. Math, actual schools

Figure A.1b. Reading, actual schools

Figure A.1c. Math, randomly assigned schools

Figure A.1d. Reading, randomly assigned schools

SGP is student growth percentile.

Note: Schools must have had at least 10 eligible students for growth estimates to be calculated, per MSDE's Every Student Succeeds Act plan. Among the 977 schools observed serving K–3 students in the statewide data provided by Maryland State Department of Education, growth estimates were calculated for 905 schools.

Source: Administrative data provided by the Maryland State Department of Education, 2014/15 to 2017/18.

Research question 1: Determining which portion of the KRA is best suited as a baseline measure in the growth model

To identify the most appropriate version of the KRA score for measuring K–3 growth, the study evaluated four different configurations of KRA scores: (1) the KRA overall score; (2) the KRA domain subscore that aligns with the SGP subject (that is, the math domain subscore for the math SGP and the reading domain subscore the for reading SGP); (3) a weighted combination of the KRA math and reading domain subscores; and (4) a weighted combination of all four KRA domain subscores. The domain weights in (3) and (4) were obtained by regressing students' grade 3 PARCC scores, separately for math and reading, on the KRA subscores. Thus, the weights reflect the extent to which each of the domain scores predict grade 3 performance. These weighted scores essentially reconfigure the KRA domain scores for the purposes of predicting grade 3 performance (as opposed to the purpose for which they were designed: assessing kindergarten readiness), potentially providing a version of the KRA score that is a better fit for the SGP model.

Four regression models were fit, separately for math and reading, on the K–3 sample. Each model corresponds with one of the four versions of the KRA score examined, as specified below:

Model 1: Overall scaled score	$PARCC3_i = \alpha + B_1 KRAOverall_i + \varepsilon_i$
Model 2: Same-subject domain score	$PARCC3_i = \alpha + B_1 KRAMath_i + \varepsilon_i$ (for the math analysis), or $PARCC3_i = \alpha + B_1 KRARead_i + \varepsilon_i$ (for the reading analysis)
Model 3: Weighted average of math and reading domain scores	$PARCC3_i = \alpha + B_1 KRAMath_i + B_2 KRARead_i + \varepsilon_i$
Model 4: Weighted average of all domain scores	$PARCC3_i = \alpha + B_1 KRAMath_i + B_2 KRARead_i + B_3 KRASocial_i + B_4 KRAPhysical_i + \varepsilon_i$

where $PARCC3_i$ refers to the grade 3 PARCC scaled score for math or reading, α refers to the constant or intercept term, $KRAOverall_i$ refers to the KRA overall scaled score, $KRAMath_i$ refers to the KRA math domain scaled subscore, $KRARead_i$ refers to the KRA reading domain scaled subscore, $KRASocial_i$ refers to the KRA social foundations domain scaled subscore, $KRAPhysical_i$ refers to the KRA physical well-being and motor development domain scaled subscore, and ε_i refers to the error term, for student i.

The regression coefficients from each model (that is, the B terms) describe how much variation in grade 3 PARCC scores is explained by the components of the KRA scores included in the model. These coefficient(s) were used as weights (along with constant term) to calculate, for each student, four predicted grade 3 PARCC scores based on each of these models. For example, model 1 predicted scores were calculated as the constant term (that is, α in model 1) plus the regression coefficient on the overall KRA score (that is, B_1 in model 1) multiplied by the student's actual overall KRA scaled score. Then, students' actual grade 3 scores were correlated with each of the four predicted scores to determine which predicted grade 3 score (and thus, which version of the KRA score) had the strongest relationship (that is, the highest correlation) with students' actual grade 3 performance.

Using the same sample to both generate and evaluate the regression coefficients, or weights, could result in unrealistically positive results for the evaluation due to "overfitting," which describes the problem that the weights were designed to fit the specific set of data from which they were generated and therefore, may not generalize well to other samples. This problem was avoided by using separate samples of students to (1) fit the regression models to derive the coefficients, and (2) evaluate the relative performance of the different versions of the KRA score (known as a "cross-validation approach"). This approach better assesses how well the results might generalize to other samples (for example, future cohorts).

Following standard practice for a "k-fold cross-validation" approach (James, Witten, Hastie, & Tibshirani, 2013), the study team split the sample in five even subsamples, excluded one subsample, and then regressed grade 3 scores, separately for math and reading, on the set of components of each of the KRA scores outlined above.[2] The regression coefficients from each of these four models were then used to predict grade 3 scores in the *excluded* subsample, and the mean squared prediction error and pairwise correlations were used to assess the similarity between the predicted grade 3 score and the actual grade 3 score in the excluded sample. The mean square error indicates how much actual scores differ from predicted scores, and the pairwise correlations describe the strength of the relationship between these two scores. The study team repeated this process with each of the five subsamples serving as the excluded sample once, and then averaged the correlations and calculated the root

[2] It is standard practice to use either 5 or 10 folds because each has been empirically shown to reduce bias and variance without being computationally intensive. The results of the analysis do not change when 10 folds are used.

mean square error (RMSE) across these folds. The correlations reported in the body of the report are the average correlations from this cross-validation approach. The results indicated by the RMSE, which describes the distance between predicted and actual scores and thus, how much the predicted value varies from the actual value, were entirely redundant with the results indicated by the correlations. Therefore, for simplicity, only the correlations are presented.

Research question 2: Calculating confidence intervals to examine the precision of schools' K–3 growth estimates, relative to schools' grades 3–4 growth estimates

To examine precision of the schools' growth estimates, the study team calculated confidence intervals around these estimates. The study team used a method known as bootstrapping, whereby schools' growth estimates were iteratively calculated using random samples of students from the original data and then examined the distribution of growth estimates calculated from those many samples. The bootstrapping approach was necessary because the method used to calculate the growth estimates (described in the preceding section) does not allow a more straightforward precision calculation.

Overview of the bootstrapping method. Bootstrapping is a method for calculating a measure of variance or precision (such as a standard error, or a confidence interval) for a sample statistic (such as a mean or a median). The standard error of any sample statistic is the standard deviation of the distribution of a large number of such statistics derived from the larger population of interest (called the sampling distribution for the statistic). For example, if one calculated the median height of a sample of women from the general population, the standard error of that median is a measure of how much that median value would differ across many different samples.[3] One can approximate the population-wide sampling distribution for a statistic by repeatedly resampling observations from the sample in the original data. In other words, instead of drawing many different samples from the population, one can draw many different samples from the original sample, with replacement (meaning that each individual observation can be selected multiple times), and then calculate the standard deviation of the statistic across those many samples. This process of resampling from the original data many times and then examining the distribution of a statistic calculated from those many samples is called bootstrapping. The number of samples is referred to as the number of bootstrap replications.

Implementing the bootstrapping method for K–3 growth estimates under full KRA administration. To calculate confidence intervals for schools' K–3 growth estimates from 2014/15 to 2017/18, when the KRA was administered to all students, each bootstrap replication consisted of the following three steps:

1. Restrict the sample to students eligible for the particular analysis (for example, the math analysis or the reading analysis) and then randomly resample students within schools, with replacement. Sampling was done within schools to ensure that the size of each school (measured as the number of students contributing information to the school's growth estimate) was constant across bootstrap replications, so that each school's confidence interval would reflect variation in which students attended the school, but not how many students attended the school. More specifically, the study team sampled students within schools using a file containing all unique student-school combinations that appeared in the original data and their associated weights. As a simplified example, imagine an original dataset containing three students, in which the first student attended school 101 for all four years from 2014/15 through 2017/18, the second student attended school 102 for all four years, and the third student attended school 101 for two years and school 102 for two years. In this example, the first step of the bootstrap replication would randomly resample students within schools 101 and 102 from a file that looks like this:

[3] Imagine repeating the exercise 1,000 times, each time drawing a different sample of women and then saving the median value calculated from that sample. The standard error of the median is equal to the standard deviation of the medians calculated for those 1,000 samples.

Student ID	School ID	Weight
1	101	1
2	102	1
3	101	0.5
3	102	0.5

2. Calculate SGP estimates as described above in the “Calculating SGP estimates” section. This step includes two noteworthy items:
    a. For the purpose of calculating these SGP estimates—which do not feed into the school growth estimates that are the focus of this study, but which are simply temporary SGP estimates that are used to calculate *confidence intervals* around schools’ growth estimates—the dataset of resampled students that came out of step 1 was restricted to the set of unique students among those sampled. In other words, students picked twice only contributed one observation to the SGP calculation, rather than two. Students who switched schools during the study period will get randomly selected more often than students who did not, because they appear more times in the file from which students are randomly resampled (in the example file above, student #3 [who switched schools during the study period] appears more often than the other students [who did not switch schools]). Thus, restricting to the set of unique students among those sampled helps ensure that more mobile students (who are likely lower-growth students) were not overweighted in the calculation of SGP estimates simply because they were more mobile than other students. Each student’s SGP estimate from this step was then attached to all of that student’s observations in the full resampled dataset, and that dataset (in which some students appeared more than once) was used in step 3 below to calculate schools’ growth estimates.
    b. Although the growth estimates produced by this study were SIMEX-adjusted to account for measurement error in students’ prior scores, the study did not make this adjustment when calculating confidence intervals around those growth estimates in order to reduce computational complexity.[4] The confidence intervals produced in this study might be slightly different from those produced using the SIMEX method. On the other hand, the width of the confidence interval around a school’s growth estimate is largely driven by school size (that is, the number of students who contribute information to the school’s growth estimate), and that sample size would not change if the SIMEX method were used. Thus, the confidence intervals produced by this study are likely a reasonable reflection of the amount of imprecision associated with schools’ growth estimates.
3. Calculate each school’s growth estimate as the weighted mean of SGP estimates across all observations for that school in the full resampled dataset (in which some students appeared more than once).

At the end of the 100 replications, for each school, the study calculated 2.5th and 97.5th percentiles of the 100 values of the school’s growth estimate.[5] These percentiles are the bootstrapped estimates of the upper and lower bounds of the 95 percent confidence interval of the growth estimate for that school. The width of the confidence interval for the school’s growth estimate equals the difference between the upper and lower bounds.

[4] In this study, a single run of the SGP package on a Windows server took 4 to 10 minutes (depending on various inputs, such as the sample size) with the SIMEX option turned off, and 7 to 15 hours with the SIMEX option turned on. Thus, taking into account the eight models the study team conducted (the model for K–3 growth, three models for grade 3–4 growth using three different cohorts, and two subjects [math and reading] for each of those four models), the team estimated that using the SIMEX option when bootstrapping would take more than 90 days, even when using multiple computer processors simultaneously.

[5] To ensure that confidence intervals could be calculated within a reasonable amount of time, this study used 100 bootstrap replications. To understand the extent to which results might change if the analysis used 1,000 replications, the study ran one of the models (the math K–3 model) using both 100 and 1,000 replications. Results were similar, differing only in the decimal places. Specifically, the average confidence interval width (in percentile points) was 12.33 for schools using 100 replications, compared to 12.49 using 1,000 replications.

As one would expect, given that medians rely on less information than means, this study found that the width of confidence intervals for schools' growth estimates was larger for median SGPs than for mean SGPs. In particular, average confidence interval widths for median SGPs (19 percentile points for math and reading) were more than 50 percent larger than those for mean SGPs (12 percentile points for math and 13 percentile points for reading).

Implementing the bootstrapping method for grades 3–4 growth estimates. To calculate confidence intervals for schools' and districts' grades 3–4 growth estimates, the study used the same procedure described above, with three changes. First, the sampling of students within schools was done using a file containing a single record for each student, in which each student was associated with his or her grade 4 school. Second, for the purpose of calculating SGP estimates in step 2 of each bootstrap replication, rather than using only one record for each unique student as was done for the K–3 growth estimates, students who were sampled repeatedly contributed their observation to the SGP calculation once for each time they were sampled. Unlike the approach used for the K–3 growth estimates, it was not necessary to restrict to unique students because each student appeared in the dataset used for sampling only once (so that the probability of a student being sampled repeatedly did not depend on whether he or she transferred between schools). Third, no weights were used in the calculation of schools' growth estimates. Rather, each student's full weight (of 1) was given to the school with which he or she was associated for grade 4, because the grade 4 assessment was designed to measure growth that occurred during grade 4, and because the majority of a student's time between the two assessments was spent in grade 4.

Research question 3: Calculating confidence intervals to examine the precision of schools' K–3 growth estimates in relation to school size

To examine precision relative to school size, the study team used the confidence intervals that were calculated for schools' K–3 growth estimates from 2014/15 to 2017/18, when the KRA was administered to all students. The procedures used to calculate these confidence intervals are described above. Schools' confidence interval widths were plotted against the schools size, which was measured as the number of students contributing to the schools' K–3 growth estimates. The study team also examined average confidence interval widths for particular quantiles of school size: 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, 99th, and 100th percentiles.

Research question 4: Implementing the bootstrapping method for K–3 growth estimates under partial KRA administration.

To calculate confidence intervals for schools' K–3 growth estimates for future years, when the KRA was administered to a subset of students, the study team used the same procedure described above, with two changes. The two changes were made to reflect that not every student who was selected to take the KRA remains in MSDE through grade 3. Because the sample of students who might contribute to a school's growth result is smaller when the KRA is administered to fewer students, further attrition among those students who were tested might have larger consequences for precision compared to when all students are administered the KRA. Specifically, the attrition will result in less precise estimates, and the two changes to the bootstrap approach under partial KRA administration are designed to ensure that this additional imprecision is measured.

First, the study team sampled from all students who took the KRA, rather than only those who also took the grade 3 PARCC. Although students who move out of the MSDE school system after taking the KRA will not receive SGPs, including these students in the sampling ensures that the resulting confidence intervals reflect imprecision that arises from this attrition. Therefore, the study team sampled students using a file containing a single record for each student with a valid KRA score, and the sampling took place within the schools where those students took the KRA. This sampling of students used the same district-specific sampling percentages that MSDE used to select a random subset of students from each classroom to whom to administer the KRA. The sampling percentages ranged from 10 to 30 percent in 2016/17, from 12 to 32 percent in 2017/18, and from 12 to 37 percent in 2018/19 (Maryland State Department of Education & Ready at Five, 2017; Maryland State Department of Education & Ready at Five, 2018; Maryland State Department of Education & Ready at Five, 2019.). For example, in 2018/19, one district selected a random subset of 21 percent of students from each classroom to take the KRA. Thus, for the 2018/19 analysis, if a particular school in that district had 350 students in the original data, the study randomly selected (350 * 0.21), or 74 students with replacement from that school. After sampling students within their KRA schools, the team then attached to those students their entire known enrollment trajectory over the four-year study period, along with the weights associated with each school they attended during that period. In other words, even though sampling was done only within students' KRA schools, the study team still used each student's entire enrollment trajectory. Although the approach sampled from a larger number of students, ultimately, the calculation of SGP estimates was restricted to sampled students who were eligible for the main K–3 analysis, meaning that it was restricted to students who had a grade 3 PARCC score and a normal grade progression between kindergarten and grade 3.

Second, when calculating SGP estimates in step 2 of each bootstrap replication, rather than using only one record for each unique student as was done for the full KRA administration, students who were sampled repeatedly contributed their observation to the SGP calculation once for each time they were sampled. Unlike the approach used for the full KRA administration, it was not necessary to restrict to unique students because each student appeared in the dataset used for sampling only once (so that the probability of a student being sampled repeatedly did not depend on whether he or she transferred between schools).

Appendix B. Supporting analyses

This appendix presents results from several supporting analyses.

The results from an analysis that compares the background characteristics and academic performance of students included in the K–3 growth estimates to students excluded from the estimates are presented in tables B.1–B.4. The analysis is presented for two groups of students. First, the analysis examines differences among all students observed in kindergarten in 2014/15, grade 1 in 2015/16, grade 2 in 2016/17, or grade 3 in 2017/18 in tables B.1 (for the math sample) and B.3 (for the reading sample). Second, the analysis examines differences among the study's population of the interest, the original cohort of students with a Kindergarten Readiness Assessment (KRA) score, in tables B.2 (for the math sample) and B.4 (for the reading sample).

Among all students observed in the grade-year combinations listed above, there are significant differences between the students included versus excluded from the analysis on nearly all of the characteristics examined (tables B.1 and B.3). Excluded students (with valid assessment score data) have slightly lower KRA scores and lower Partnership for Assessment of Readiness for College and Career (PARCC) scores. They are more likely to be English learners, eligible for special education services, eligible for free or reduced-price lunch, and in a Title I school. In addition, excluded students are less likely to be female and White. Most of these differences are greater than 0.05 standard deviations.

Among the original cohort of students who completed the KRA in 2014/15 (tables B.2 and B.4)—which is the population of interest for this study—the differences between the students included versus excluded from the analysis are similar to those among all students. However, in terms of standard deviations, the differences are larger for special education status and smaller (less than 0.05 standard deviations) for English learner status and White. The model does not implement methods to account for these differences because face validity is important for school performance measures, and such adjustments can erode stakeholders' confidence in the measure. For example, one approach would be to use nonresponse weights that assign larger weights to students who are underrepresented in the sample, so the sample more closely reflects the original cohort of students who took the KRA. However, it is difficult for schools, teachers, and parents to understand why the performance of particular students should receive relatively more weight.

Table B.1. Characteristics of students included in and excluded from the K–3 math student growth percentile analysis

Characteristic	Mean value for students included in the analysis	Mean value for students excluded from the analysis[a]	Difference in mean value for students included and excluded from the analysis (in standard deviation units)
2014/15 KRA overall z-score	0.06	-0.33*[b]	0.39
2014/15 KRA overall scale score	269.64	264.69*[b]	0.39
Emerging KRA readiness (percentage)	15.27	29.67*[b]	-0.38
Approaching KRA readiness (percentage)	36.43	32.36*[b]	0.08
Demonstrating KRA readiness (percentage)	48.3	37.97*[b]	0.21
2017/18 PARCC z-score	0.03	-0.13*[c]	0.16
2017/18 PARCC scale score	742.65	736.3*[c]	0.16
In a Title I school (percentage)	36.25	39.32*	-0.06
Female (percentage)	49.37	47.38*	0.04
Race/ethnicity (percentage)			
African American	32.41	35.79*	-0.07
American Indian/Alaska Native	0.26	0.26	0
Asian	5.91	7.81*	-0.08
Hawaiian/Pacific Islander	0.13	0.28*	-0.04
White	38.34	31.08*	0.15
Multiple races	5.18	5.04	0.01
Hispanic	17.78	19.74*	-0.05
Migrant (percentage)	0	0.01	-0.01
English learner (percentage)	11.43	19.74*	-0.24
Special education (percentage)	10.98	16.09*	-0.15
Had one or more disciplinary infractions (percentage)	1.96	2.22*	-0.02
Number of disciplinary infractions	0.03	0.04*	-0.03
Eligible for free or reduced-price lunch (percentage)	46.14	50.02*	-0.08
Number of students	**54,393**	**26,467**	

KRA is Kindergarten Readiness Assessment. PARCC is Partnership for Assessment of Readiness for College and Careers.

* Significantly different from students included in the analysis at the .05 level, two-tailed test.

Note: This table shows the mean value of each characteristic in the 2017/18 school year (or the latest year available for a given student).

a. These are students who (1) have no 2014/15 KRA score (for example, because they entered the Maryland public school system after the 2014/15 KRA was administered); (2) have no 2017/18 assessment score (for example, because they left the Maryland public school system before the 2017/18 assessment was administered); (3) have a 2017/18 score but the test code associated with that score was not grade 3; (4) had a non-normal grade progression at any point from 2014/15 through 2017/18, or (5) have a 2017/18 alternate assessment score, but not a 2017/18 PARCC score.

b. Among excluded students with a valid KRA score (n = 9,125; about 34 percent of excluded students).

c. Among excluded students with a valid PARCC score (n = 13,499; about 51 percent of excluded students).

Source: Administrative data provided by the Maryland State Department of Education.

Table B.2. Characteristics of students included in and excluded from the K–3 math student growth percentile analysis, among students with a Kindergarten Readiness Assessment (KRA) score

Characteristic	Mean value for students included in the analysis	Mean value for students excluded from the analysis[a]	Difference in mean value for students included and excluded from the analysis (in standard deviation units)
2014/15 KRA overall z-score	0.06	-0.33*	0.39
2014/15 KRA overall scale score	269.64	264.69*	0.39
Emerging KRA readiness (percentage)	15.27	29.67*	-0.38
Approaching KRA readiness (percentage)	36.43	32.36*	0.08
Demonstrating KRA readiness (percentage)	48.3	37.97*	0.21
In a Title I school (percentage)	36.25	38.81*	-0.05
Female (percentage)	49.37	45.92*	0.07
Race/ethnicity (percentage)			
African American	32.41	34.09*	-0.04
American Indian/Alaska Native	0.26	0.28	-0.01
Asian	5.91	5.92	0
Hawaiian/Pacific Islander	0.13	0.28*	-0.04
White	38.34	39.29	-0.02
Multiple races	5.18	5.74*	-0.03
Hispanic	17.78	14.39*	0.09
Migrant (percentage)	0	0	0
English learner (percentage)	11.43	12.51*	-0.03
Special education (percentage)	10.98	19.22*	-0.25
Had one or more disciplinary infractions (percentage)	1.96	2.07	-0.01
Number of disciplinary infractions	0.03	0.04*	-0.03
Eligible for free or reduced-price lunch (percentage)	46.14	49.63*	-0.07
Number of students	**54,393**	**9,125**	

PARCC is Partnership for Assessment of Readiness for College and Careers.

* Significantly different from students included in the analysis at the .05 level, two-tailed test.

Note: This table shows the mean value of each characteristic in the 2017/18 school year (or the latest year available for a given student).

a. These are students who (1) have no 2017/18 test score (for example, because they left the Maryland public school system before the 2017/18 test was administered); (2) have a 2017/18 score but the test code associated with that score was not grade 3; (3) had a non-normal grade progression at any point from 2014/15 through 2017/18; or (4) have a 2017/18 alternate test score, but not a 2017/18 PARCC score.

Source: Administrative data provided by the Maryland State Department of Education.

Table B.3. Characteristics of students included in and excluded from the K–3 reading student growth percentile analysis

Characteristic	Mean value for students included in the analysis	Mean value for students excluded from the analysis[a]	Difference in mean value for students included and excluded from the analysis (in standard deviation units)
2014/15 KRA overall z-score	0.06	-0.33* [b]	0.39
2014/15 KRA overall scale score	269.64	264.68* [b]	0.39
Emerging KRA readiness (percentage)	15.26	29.72* [b]	-0.38
Approaching KRA readiness (percentage)	36.43	32.32* [b]	0.09
Demonstrating KRA readiness (percentage)	48.31	37.96* [b]	0.21
2017/18 PARCC z-score	0.03	-0.13* [c]	0.16
2017/18 PARCC scale score	737.44	730.31* [c]	0.16
In a Title I school (percentage)	36.25	39.33*	-0.06
Female (percentage)	49.37	47.38*	0.04
Race/ethnicity (percentage)			
African American	32.41	35.78*	-0.07
American Indian/Alaska Native	0.26	0.26	0
Asian	5.91	7.82*	-0.08
Hawaiian/Pacific Islander	0.13	0.28*	-0.04
White	38.34	31.08*	0.15
Multiple races	5.17	5.04	0.01
Hispanic	17.78	19.74*	-0.05
Migrant (percentage)	0	0.01	-0.01
English learner (percentage)	11.43	19.75*	-0.24
Special education (percentage)	10.97	16.1*	-0.15
Had one or more disciplinary infractions (percentage)	1.96	2.23*	-0.02
Number of disciplinary infractions	0.03	0.04*	-0.03
Eligible for free or reduced-price lunch (percentage)	46.13	50.04*	-0.08
Number of students	**54,397**	**26,463**	

KRA is Kindergarten Readiness Assessment. PARCC is Partnership for Assessment of Readiness for College and Careers.

* Significantly different from students included in the analysis at the .05 level, two-tailed test.

Note: This table shows the mean value of each characteristic in the 2017/18 school year (or the latest year available for a given student).

a. These are students who (1) have no 2014/15 KRA score (for example, because they entered the Maryland public school system after the 2014/15 KRA was administered); (2) have no 2017/18 assessment score (for example, because they left the Maryland public school system before the 2017/18 assessment was administered); (3) have a 2017/18 score but the test code associated with that score was not grade 3; (4) had a non-normal grade progression at any point from 2014/15 through 2017/18, or (5) have a 2017/18 alternate assessment score, but not a 2017/18 PARCC score.

b. Among excluded students with a valid KRA score (*n* = 9,115; about 34 percent of excluded students).

c. Among excluded students with a valid PARCC score (*n* = 13,320; about 50 percent of excluded students).

Source: Administrative data provided by the Maryland State Department of Education.

Table B.4. Characteristics of students included in and excluded from the K–3 reading student growth percentile analysis, among students with a Kindergarten Readiness Assessment (KRA) score

Characteristic	Mean value for students included in the analysis	Mean value for students excluded from the analysis[a]	Difference in mean value for students included and excluded from the analysis (in standard deviation units)
2014/15 KRA overall z-score	0.06	-0.33*	0.39
2014/15 KRA overall scale score	269.64	264.68*	0.39
Emerging KRA readiness (percentage)	15.26	29.72*	-0.38
Approaching KRA readiness (percentage)	36.43	32.32*	0.09
Demonstrating KRA readiness (percentage)	48.31	37.96*	0.21
In a Title I school (percentage)	36.25	38.81*	-0.05
Female (percentage)	49.37	45.92*	0.07
Race/ethnicity (percentage)			
African American	32.41	34.08*	-0.04
American Indian/Alaska Native	0.26	0.29	-0.01
Asian	5.91	5.93	0
Hawaiian/Pacific Islander	0.13	0.29*	-0.04
White	38.34	39.29	-0.02
Multiple races	5.17	5.74*	-0.03
Hispanic	17.78	14.38*	0.09
Migrant (percentage)	0	0	0
English learner (percentage)	11.43	12.52*	-0.03
Special education (percentage)	10.97	19.25*	-0.25
Had one or more disciplinary infractions (percentage)	1.96	2.09	-0.01
Number of disciplinary infractions	0.03	0.04*	-0.03
Eligible for free or reduced-price lunch (percentage)	46.13	49.69*	-0.07
Number of students	**54,397**	**9,121**	

PARCC is Partnership for Assessment of Readiness for College and Careers.

* Significantly different from students included in the analysis at the .05 level, two-tailed test.

Note: This table shows the mean value of each characteristic in the 2017/18 school year (or the latest year available for a given student).

a. These are students who (1) have no 2017/18 test score (for example, because they left the Maryland public school system before the 2017/18 test was administered); (2) have a 2017/18 score but the test code associated with that score was not grade 3; (3) had a non-normal grade progression at any point from 2014/15 through 2017/18; or (4) have a 2017/18 alternate test score, but not a 2017/18 PARCC score.

Source: Administrative data provided by the Maryland State Department of Education.

The results from an analysis that compares the background characteristics and academic performance of students included in the K–3 growth estimates who do and do not move schools within school years are presented in table B.5 (results for the math sample) and table B.6 (results for the reading sample). Twenty-four percent of the students eligible for the K–3 analysis moved schools within one or more school years. For those students, on average, the enrollment data cover 89 percent of the period from kindergarten to grade 3, meaning that the study does not observe where those students attended school for only 11 percent of the study period. There are significant differences between the students who move and do not move within school years on nearly all of the characteristics examined, and most of these differences are greater than 0.05 standard deviations.

Table B.5. Characteristics of students who do and do not move schools within school years, K–3 math student growth percentile (SGP) analysis

Characteristic	Mean value for students who move schools within school years	Mean value for students who do not move schools within school years	Difference in mean value for students who do and don’t move schools within school years (in standard deviation units)
2014/15 KRA overall z-score	-0.09*	0.1	-0.19
2014/15 KRA overall scale score	267.78*	270.24	-0.19
Emerging KRA readiness (percentage)	19.39*	13.94	0.14
Approaching KRA readiness (percentage)	38.09*	35.89	0.05
Demonstrating KRA readiness (percentage)	42.52*	50.17	-0.15
2017/18 PARCC z-score	-0.17*	0.1	-0.27
2017/18 PARCC scale score	734.7*	745.22	-0.27
Math SGP estimate	46.25*	51.212	-0.17
Reading SGP estimate	46.94*	51.01	-0.14
In a Title I school (percentage)	45.47*	33.28	0.25
Female (percentage)	49.3	49.39	0
Race/ethnicity (percentage)			
African American	44.06*	28.65	0.33
American Indian/Alaska Native	0.16*	0.29	-0.03
Asian	6.82*	5.62	0.05
Hawaiian/Pacific Islander	0.13	0.13	0
White	24.21*	42.9	-0.39
Multiple races	4.96	5.25	-0.01
Hispanic	19.67*	17.17	0.06
Migrant (percentage)	0.01	0	0.01
English learner (percentage)	14.11*	10.57	0.1
Special education (percentage)	12.07*	10.62	0.04
Had one or more disciplinary infractions (percentage)	0.05*	0.03	0.09
Number of disciplinary infractions	3.09*	1.6	0.11
Eligible for free or reduced-price lunch (percentage)	59.98*	41.67	0.37
Number of students	**13,269**	**41,124**	

KRA is Kindergarten Readiness Assessment. PARCC is Partnership for Assessment of Readiness for College and Careers.

* Significantly different from students who do not move schools within school years at the .05 level, two-tailed test.

Note: This table shows the mean value of each characteristic in the 2017/18 school year (or the latest year available for a given student).

Source: Administrative data provided by the Maryland State Department of Education.

Table B.6. Characteristics of students who do and do not move schools within school years, K–3 reading student growth percentile (SGP) analysis

Characteristic	Mean value for students who move schools within school years	Mean value for students who do not move schools within school years	Difference in mean value for students who do and don't move schools within school years (in standard deviation units)
2014/15 KRA overall z-score	-0.09*	0.1	-0.19
2014/15 KRA overall scale score	267.78*	270.24	-0.19
Emerging KRA readiness (percentage)	19.38*	13.93	0.14
Approaching KRA readiness (percentage)	38.11*	35.89	0.05
Demonstrating KRA readiness (percentage)	42.51*	50.18	-0.15
2017/18 PARCC z-score	-0.16*	0.09	-0.25
2017/18 PARCC scale score	729.28*	740.08	-0.25
Math SGP estimate	46.29*	51.22	-0.17
Reading SGP estimate	46.9*	51	-0.14
In a Title I school (percentage)	45.45*	33.28	0.25
Female (percentage)	49.31	49.39	0
Race/ethnicity (percentage)			
African American	44.06*	28.65	0.33
American Indian/Alaska Native	0.16*	0.29	-0.03
Asian	6.82*	5.62	0.05
Hawaiian/Pacific Islander	0.13	0.13	0
White	24.22*	42.9	-0.39
Multiple races	4.96	5.25	-0.01
Hispanic	19.66*	17.17	0.06
Migrant (percentage)	0.01	0	0.01
English learner (percentage)	14.1*	10.57	0.1
Special education (percentage)	12.09*	10.61	0.04
Had one or more disciplinary infractions (percentage)	0.05*	0.03	0.09
Number of disciplinary infractions	3.09*	1.59	0.11
Eligible for free or reduced-price lunch (percentage)	59.98*	41.66	0.37
Number of students	**13,276**	**41,121**	

KRA is Kindergarten Readiness Assessment. PARCC is Partnership for Assessment of Readiness for College and Careers.
*Significantly different from students who do not move schools within school years at the .05 level, two-tailed test.
Note: This table shows the mean value of each characteristic in the 2017/18 school year (or the latest year available for a given student).
Source: Administrative data provided by the Maryland State Department of Education.

The relationship between schools' characteristics and the percentage of the schools' students who were excluded from the analyses, among all students who were observed as enrolled in the school during the K–3 period, is examined in Table B.7. The results indicate that the percentage of a schools' students excluded from the analyses (largely due to students entering the school system after the KRA or leaving the school system before the grade 3 PARCC assessment; see table A.1) is related to the schools' growth estimate and the academic performance of the students it serves. There is a weak correlation between schools' K–3 growth estimates and the percentage of students who were excluded from the analyses (-0.2), but the coefficient from a regression of the growth estimate on the percentage of students excluded from the analyses indicate that a 1 percentage point increase in the schools' students who were excluded from the analysis is associated with a decrease in the schools' K–3 growth estimates of 0.3 (for math and reading), and these results are significant. There are moderate correlations between the percentage of students excluded from the analysis and the schools' average KRA and grade 3 PARCC math and reading scores, among all students with a score (-0.7, -0.5, -0.5, respectively), and the regression coefficients (-0.3, -0.7, -0.6, respectively) are significant. Summary statistics for schools' K–3 growth estimates are presented in Table B.8. Schools' growth estimates range from 11 to 93 for math, and 12 to 91 for reading.

Table B.7. Relationship between schools' characteristics and the percentage of students in the schools' sample who were excluded from the analyses

Characteristic	Correlation with the percentage of students in the schools' sample who were excluded from the analyses	Coefficient from a regression of the characteristic on the percentage of students in the schools' sample who were excluded from the analyses
Average 2014/15 KRA overall scale score[a]	-0.66	-0.29*
Average 2017/18 PARCC math scale score[a]	-0.54	-0.68*
Average 2017/18 PARCC reading scale score[a]	-0.50	-0.64*
K–3 math growth estimate	-0.22	-0.32*
K–3 reading growth estimate	-0.18	-0.25*
Indicator for being a Title I school	-0.10	-0.23*
Female (percentage of students)	-0.42	-0.22*
Race/ethnicity (percentage of students)		
African American	0.16	0.24*
American Indian/Alaska Native	-0.05	0.00
Asian	-0.10	-0.04*
Hawaiian/Pacific Islander	-0.06	0.00
White	-0.11	-0.16*
Multiple races	-0.13	-0.03*
Hispanic	-0.02	-0.01
Migrant (percentage of students)	-0.03	0.00
English learner (percentage of students)	-0.08	-0.06**
Special education (percentage of students)	0.85	0.89*
Had one or more disciplinary infractions (percentage of students)	0.12	0.03*
Average number of yearly disciplinary infractions	0.14	0.00*
Eligible for free or reduced-price lunch (percentage of students)	-0.03	-0.04

KRA is Kindergarten Readiness Assessment. PARCC is Partnership for Assessment of Readiness for College and Careers.

*/** Significantly different from zero at the .05/.01 level, two-tailed test.

Note: Each school's sample includes all Maryland students observed as enrolled in the school in kindergarten in 2014/15, grade 1 in 2015/16, grade 2 in 2016/17, or grade 3 in 2017/18. Across all schools, the 50th and 90th percentiles of the percentage of students excluded from the analyses were 20 and 33, respectively.

a. Average among all students with a score, whether or not they were excluded from the K-3 growth analyses.

Source: Administrative data provided by the Maryland State Department of Education, 2014/15 to 2017/18.

Table B.8. Mean, standard deviation, and percentiles of schools' K–3 growth estimates

Statistic	Schools' K–3 growth estimates Math	Schools' K–3 growth estimates Reading
Mean	48[a]	48[a]
Standard deviation	14	14
Minimum value	11	12
1st percentile	16	16
5th percentile	21	22
10th percentile	27	29
25th percentile	39	41
50th percentile (median)	50	49
75th percentile	58	58
90th percentile	66	65
95th percentile	70	69
99th percentile	77	76
Maximum value	93	91

Note: See appendix A for details on how growth estimates were calculated. Schools must have had at least 10 eligible students for growth estimates to be calculated, per MSDE's Every Student Succeeds Act plan. Among the 977 schools observed serving K–3 students in the statewide data provided by Maryland State Department of Education, growth estimates were calculated for 905 schools.

a. The mean value is not 50 because it is a simple mean of schools' growth estimates, as opposed to a weighted mean using school size as the weight, and because larger schools tend to have slightly higher growth estimates than smaller schools.
Source: Administrative data provided by the Maryland State Department of Education, 2014/15 to 2017/18.

The correlations between schools' K–3 growth estimates and grades 3-4 growth estimates are presented in Table B.9. For all analyses that compared K–3 growth estimates to grades 3–4 growth estimates, results were similar regardless of which grades 3–4 cohort was used in the analysis. The cohort of students who were in grade 3 in 2016/17 and grade 4 in 2017/18 are considered the primary cohort for analysis for two reasons: (1) they are the closest in age to the cohort for the K–3 growth analysis (they are one year older), and (2) they took the PARCC in the same year that the K–3 cohort did (2017/18), so any changes to the PARCC in that year should not affect the comparison of the K–3 results with the grades 3–4 results.

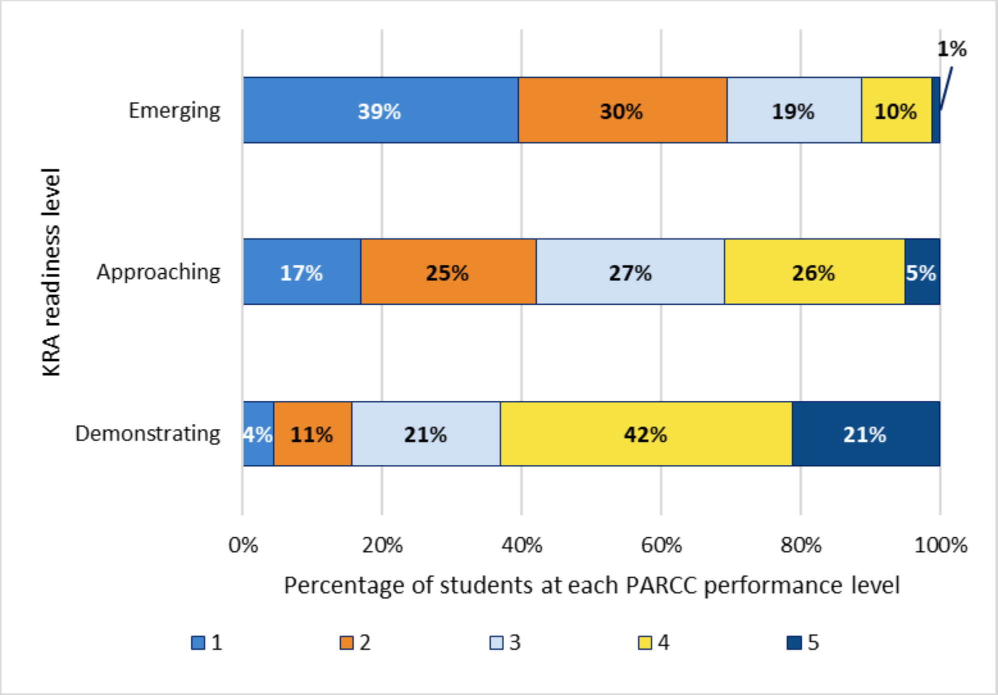
Table B.9. Correlations between schools' K–3 growth estimates and grades 3–4 growth estimates

Grades and school years	Math	Reading
3–4; 2014/15 to 2015/16	0.45	0.52
3–4; 2015/16 to 2016/17	0.37	0.41
3–4; 2016/17 to 2017/18	0.40	0.37

Note: See appendix A for details on how growth estimates were calculated.
Source: Administrative data provided by the Maryland State Department of Education, 2014/15 to 2017/18.

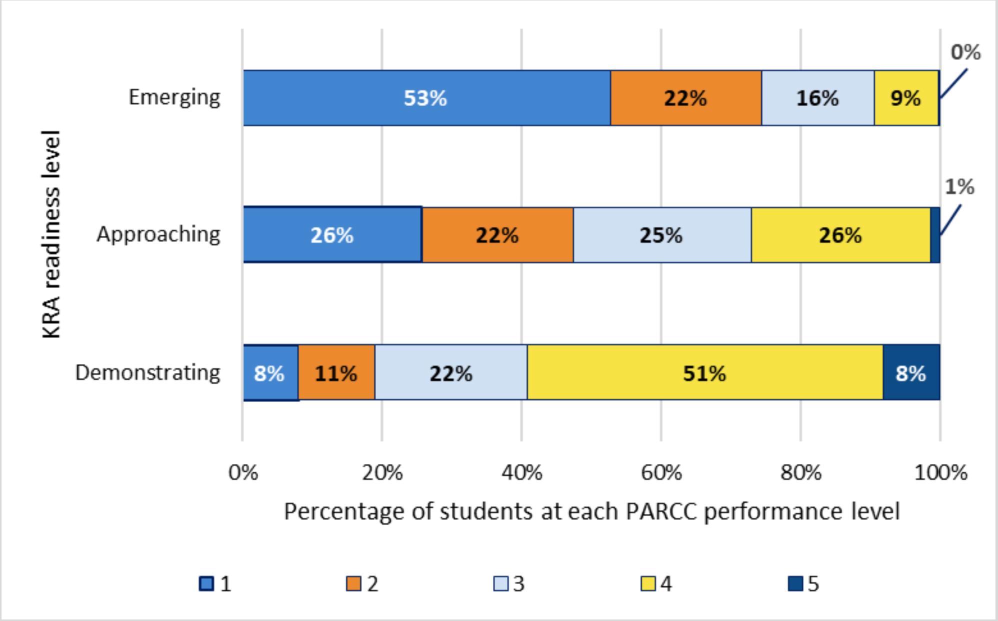
The percentages of students at each PARCC performance level (for math and reading), by KRA readiness level, are shown in Figures B.1 and B.2. Students with high KRA scores tended to also have high grade 3 PARCC scores, and students with low KRA scores tended to have low grade 3 PARCC scores. For example, 88 percent of students who had "emerging" math skills on the KRA did not meet expectations on the grade 3 PARCC, versus 36 percent of students who were "demonstrating" math skills.

Figure B.1. Percentage of students at each grade 3 Partnership for Assessment of Readiness for College and Careers (PARCC) math performance level, by Kindergarten Readiness Assessment (KRA) readiness level



Note: PARCC performance levels are defined as: 1=Did Not Yet Meet Expectations, 2=Partially Met Expectations; 3=Approached Expectations; 4=Met Expectations; and 5=Exceeded Expectations.
Source: Administrative data provided by the Maryland State Department of Education, 2014/15 to 2017/18.

Figure B.2. Percentage of students at each grade 3 Partnership for Assessment of Readiness for College and Careers (PARCC) reading performance level, by Kindergarten Readiness Assessment (KRA) readiness level



Note: PARCC performance levels are defined as: 1=Did Not Yet Meet Expectations, 2=Partially Met Expectations; 3=Approached Expectations; 4=Met Expectations; and 5=Exceeded Expectations.
Source: Administrative data provided by the Maryland State Department of Education, 2014/15 to 2017/18.

Confidence interval widths for schools' K–3 growth estimates are plotted against school size in Figure B.3. Each dot represents a school. The precision of schools' estimates decreases dramatically below a certain school size, as indicated by the steep downward slope of the data for schools with fewere than 50 students. Additionally, there are diminishing returns to precision (that is, each additional student adds less precision than the last), as indicated by the fact that precision levels out around 100 students; schools with more than 150 students have growth estimates that are about as precise as schools with only 100 students.

Figure B.3. Confidence interval widths for schools' growth estimates plotted against school size, by subject
Percentile points

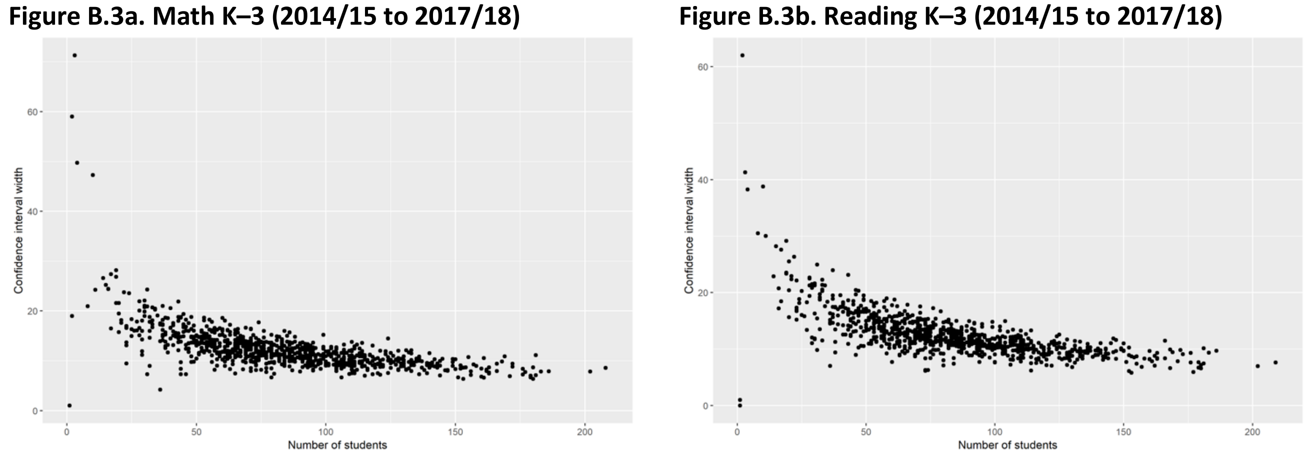


Figure B.3a. Math K–3 (2014/15 to 2017/18)

Figure B.3b. Reading K–3 (2014/15 to 2017/18)

Note: These figures show the relationship between school size (measured as the number of students contributing information to the school's growth estimate) and the width of the confidence interval around schools' growth estimates. Each dot represents a school. See appendix A for details on how confidence intervals were calculated.
Source: Administrative data provided by the Maryland State Department of Education.

The distributions of average confidence interval widths, by subject and the percentage of students who take the KRA, are shown in Figure B.4. The results for math using the 2016/17 sampling percentages were presented and discussed in the body of the report. This figure shows that these results are similar to the results for reading using the 2016/17 sampling percentages, and to the results for both subjects using the 2017/18 and 2018/19 sampling percentages (which is expected given that the percentage of students taking the KRA is similar over these three years).

Figure B.4. Distribution of average confidence interval widths for schools' K–3 growth estimates, by subject and percentage of students who take the Kindergarten Readiness Assessment (KRA)

Percentile points

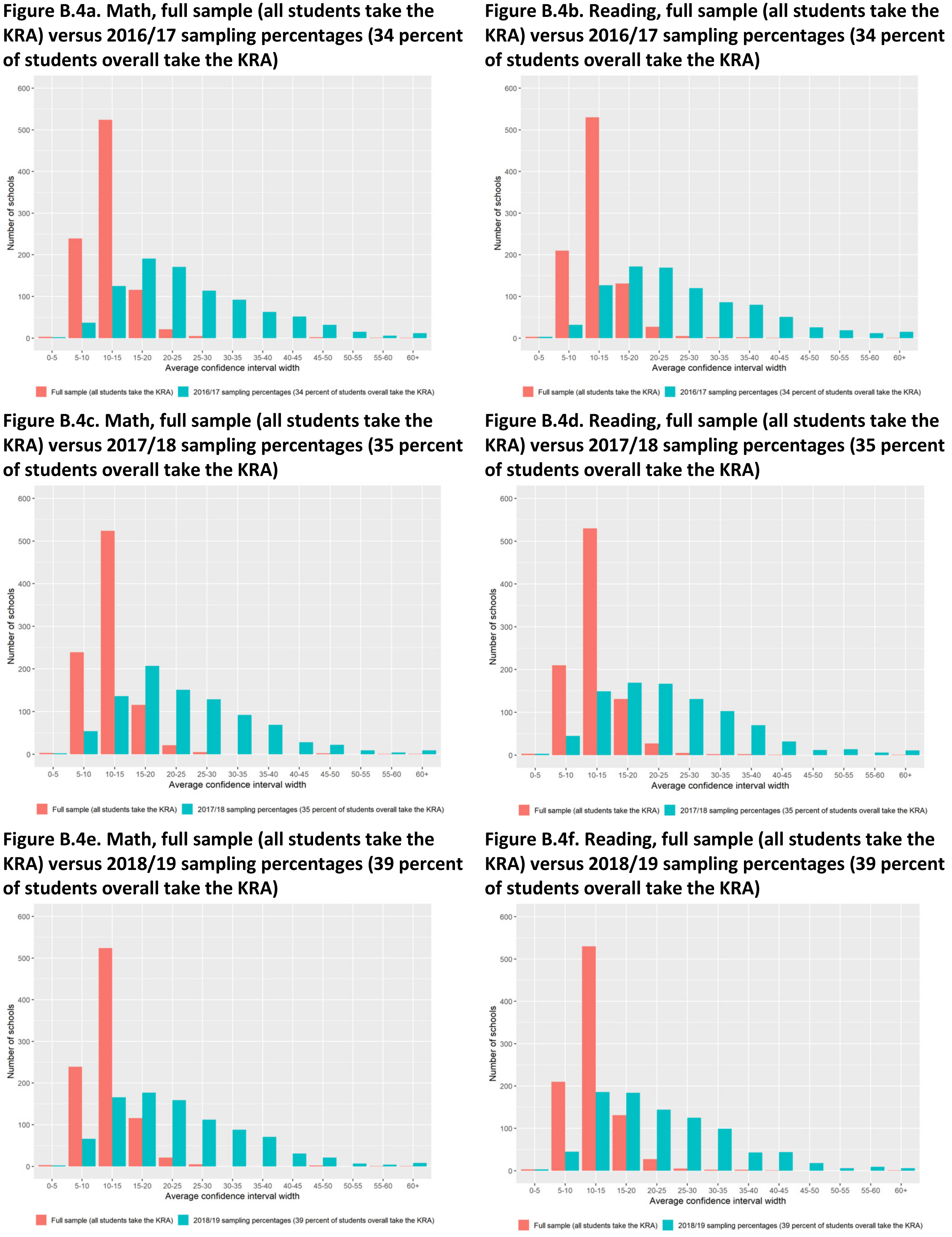


Figure B.4a. Math, full sample (all students take the KRA) versus 2016/17 sampling percentages (34 percent of students overall take the KRA)

Figure B.4b. Reading, full sample (all students take the KRA) versus 2016/17 sampling percentages (34 percent of students overall take the KRA)

Figure B.4c. Math, full sample (all students take the KRA) versus 2017/18 sampling percentages (35 percent of students overall take the KRA)

Figure B.4d. Reading, full sample (all students take the KRA) versus 2017/18 sampling percentages (35 percent of students overall take the KRA)

Figure B.4e. Math, full sample (all students take the KRA) versus 2018/19 sampling percentages (39 percent of students overall take the KRA)

Figure B.4f. Reading, full sample (all students take the KRA) versus 2018/19 sampling percentages (39 percent of students overall take the KRA)

Note: See appendix A for details on how confidence intervals were calculated.

Source: Administrative data provided by the Maryland State Department of Education, 2014/15 to 2017/18.